Basic statistical concepts and figures in R

Dr. Yusheng Zhao

Department of breeding research,

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)

(Email: zhao@ipk-gatersleben.de)

Instruction structure

- **Descriptive and inferential statistics**
- **Statistical testing**
- **Liner regression and ANOVA**

Descriptive and inferential statistics

Statistics:

Average, variance, standard deviation, median, quartile, frequency distribution, correlation

Figures:

boxplot, histogram, QQ-plot, barplot, pie chart, scatterplot

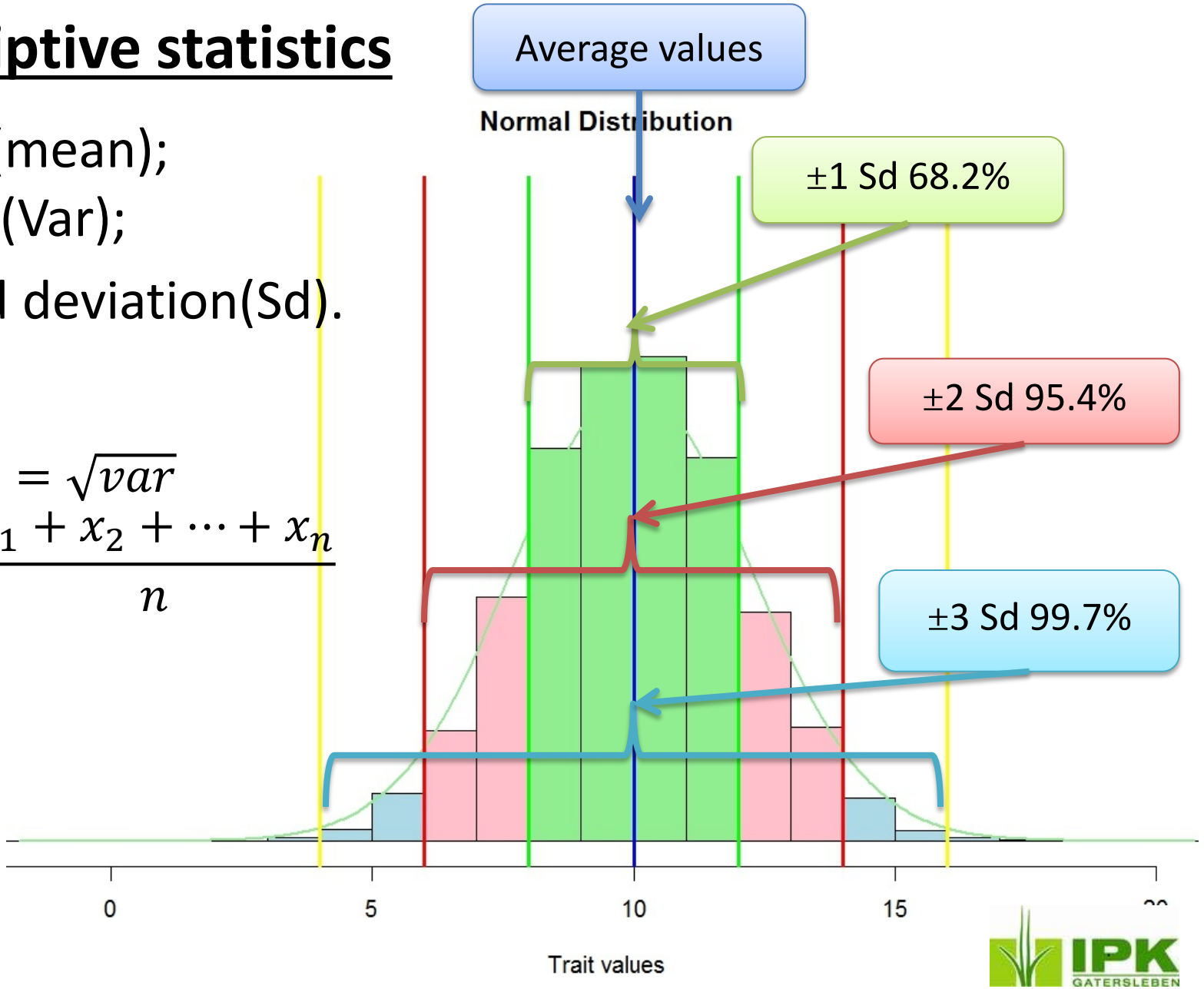
IPK GATERSLEBEN

Descriptive statistics

Average(mean);
Variance(Var);
Standard deviation(Sd).

$$Sd = \sqrt{var}$$

$$Mean = \frac{x_1 + x_2 + \cdots + x_n}{n}$$



Example data

A wheat data set of 80 genotypes. Plants in green house under control condition.

12 traits: Grain yield(GY), Seed yield(SY), Thousand kernel weight(TKW), Seed width(SDW), Seed length(SDL), Stem length(STM),Spike length(SPK), Thousand kernel weight of main spike(TKW_M), Seed width of main spike(SDW_M),Seed length of main spike(SDL_M), Tiller number per plant(Tiller),Leaf number per stem(Leaf)

	Country	GY	SY	TKW	SDW	SDL	STM	SPK	TKW_M	SDW_M	SDL_M	Tiller	Leaf
Gen_1	Germany	5.1	130	39.6	3.5	6	62.1	11.3	45.3	4.2	6.4	3	4
Gen_2	Germany	5.7	168	34.2	3.6	6.6	68.7	12.4	44	4.2	6.3	11	3
Gen_3	Germany	4.8	182	26.6	3	6.5	47.4	14	32.8	3.8	6.5	4	3
Gen_4	…	…	…	…	…	…	…	…	…	…	…	…	…

IPK GATERSLEBEN

Descriptive statistics

1. Average values

let $x = (x_1, x_2, x_3, \dots, x_n)$, average values of x is:

$$\bar{x} = \frac{x_1, + x_2 + x_3 + \dots + x_n}{n} = \frac{1}{n}\sum_1^n x_i$$

Use R function mean()

```
> setwd( choose.dir() )  ## set the work dir
> Data<-read.table("Data.txt",header=TRUE)
> head(Data)
Country  GY  SY  TKW SDW SDL  STM  SPK TKW_
Gen_1 Germany 5.1 130 39.6 3.5 6.0 62.1 11.3
Gen_2 Germany 5.7 168 34.2 3.6 6.6 68.7 12.4
… … … … … … … … … … … … … …
> mean(Data$GY, na.rm=TRUE)
[1] 4.739437
```

"na.rm=TRUE" will ignore missing values

This is the target trait: grain yield "GY"

Descriptive statistics

1. Average values

Use R function apply() to get average value for all traits

```
## define a new function
> Mean.rm<-function(x){
+ a<-mean(x,na.rm=TRUE)
+ }
> apply(Data[,2:12], 2, Mean.rm)
```

```
      GY        SY       TKW       SDW       SDL       STM       SPK    TKW_M
 4.739437 136.347222  34.223611  3.484932  6.335616 63.892500 12.091250 42.126027
   SDW_M     SDL_M     Tiller
 3.790278  6.621918  4.337500
```

The first column is not traits

This parameter is set as “2” in order to apply the function “Mean.rm” to all the columns of data,
If this is equal to ”1”, then apply “Mean.rm” to all the rows

Descriptive statistics

2. Variance and standard deviation

let $x = (x_1, x_2, x_3, \dots, x_n)$,

then:

$$Var(x) = \frac{\sum_i^n (x_i - \bar{x})^2}{n-1}, \; Sd = \sqrt{Var(x)}$$

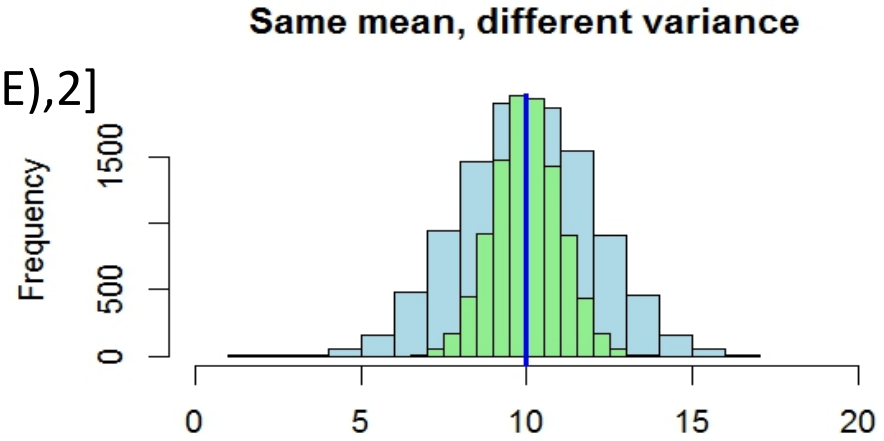
R function var(), sd()

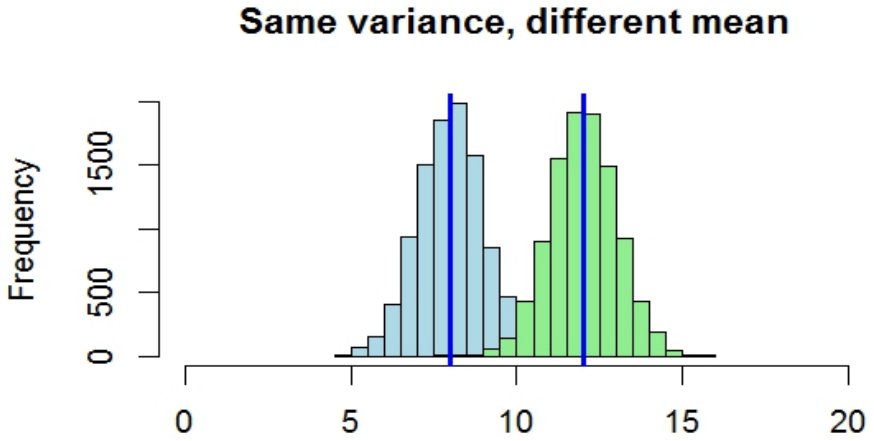
```
> ## variance and sd
> var(Data[,2],na.rm=TRUE)
[1] 3.108423
> sd(Data[,2],na.rm=TRUE)
[1] 1.763072
```

IPK GATERSLEBEN

Descriptive statistics

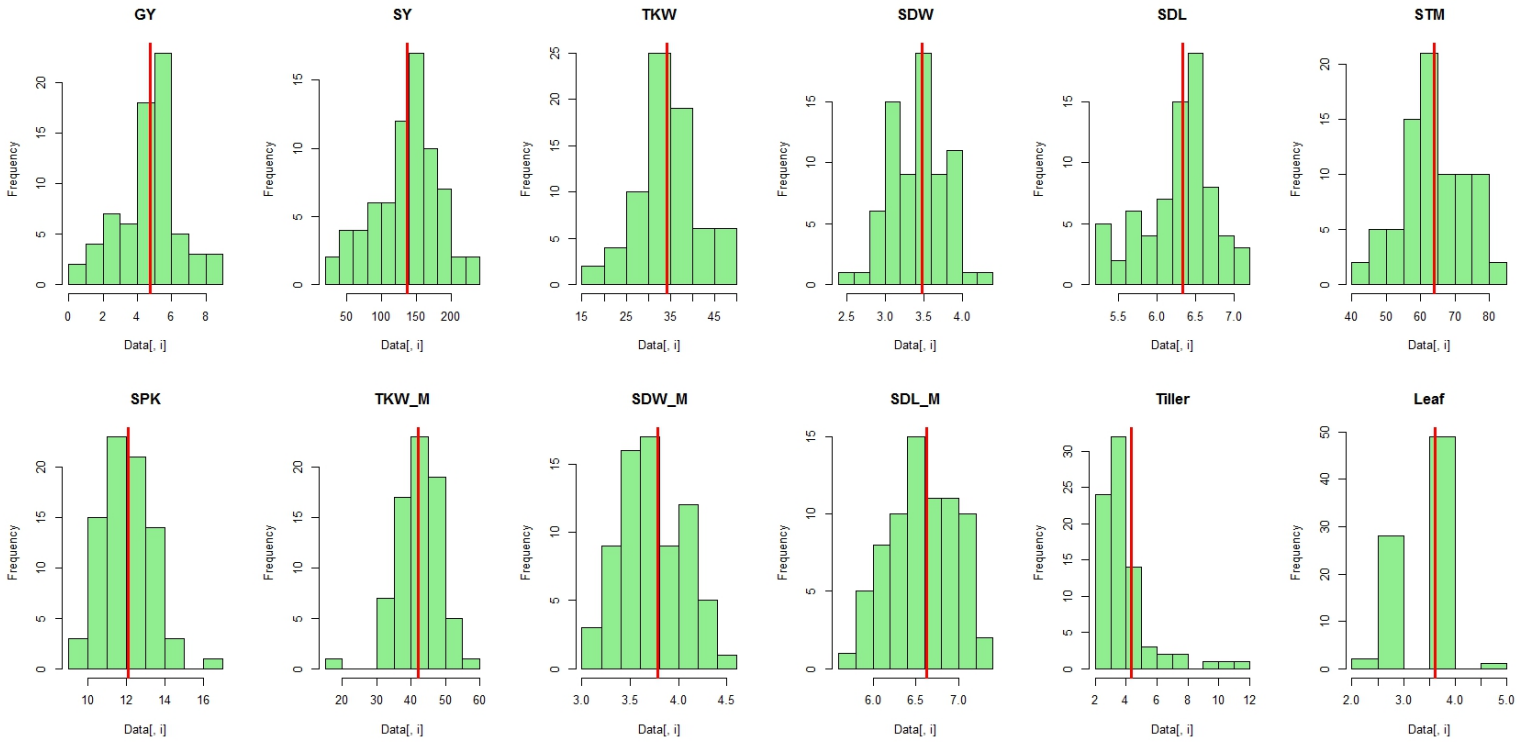
```
> VVV<-Data[-which(is.na(Data[,2])==TRUE),2]
> var(VVV)
[1] 3.108423
> sqrt(var(VVV))
[1] 1.763072
> sum((VVV-mean(VVV))^2)/(length(VVV)-1)
[1] 3.108423
```





Descriptive statistics

```
## Histogram show mean and sd
layout(matrix(c(1:12),2,6,byrow=TRUE))
for(i in 2:13){
hist(Data[,i],col="lightgreen",main=paste(colnames(Data)[i]))
abline(v=mean(Data[,i],na.rm=TRUE),lwd=3,col="red")
}
```



Descriptive statistics

Five Number Summary:

The summary of data consists of minimum values, the **1st** quartile(25%), median(50%), the 3rd quartile(75%), and maximum values.

R function max(), min(), range(), quantile()

```
> max(Data[,2],na.rm=TRUE)
[1] 8.7
>  min(Data[,2],na.rm=TRUE)
[1] 0.4
>  range(Data[,2],na.rm=TRUE)
[1] 0.4 8.7
> quantile(Data[,2],na.rm=TRUE)
  0%  25%  50%  75% 100%
 0.4  4.0  5.0  5.7  8.7
```

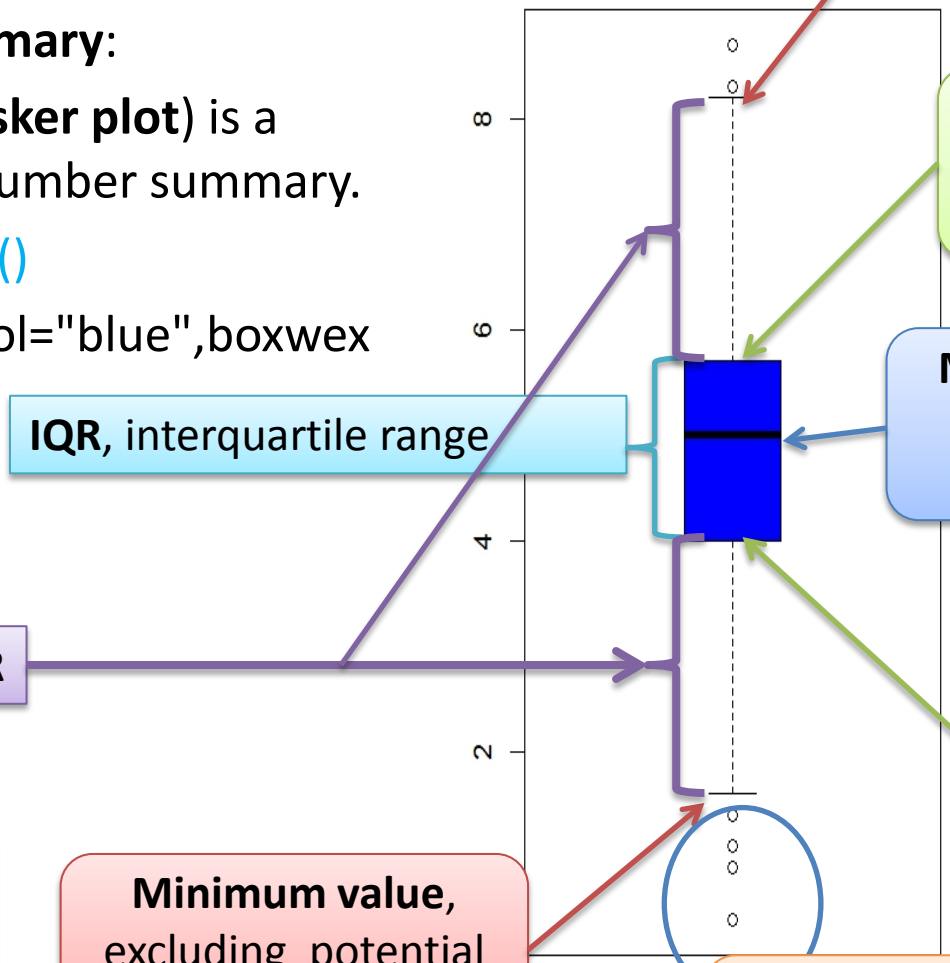
Descriptive statistics

Five Number Summary:

Box plot (**box-whisker plot**) is a graph of the five number summary.

R function boxplot()

boxplot(Data[,2],col="blue",boxwex =0.5,lwd=1.5)



Maximum value, excluding potential outliers

the 3rd quartile, 75% of data is smaller than this value

Median, 50% of data is larger or smaller than this value

IQR, interquartile range

1.5 time IQR

The 1st quartile, 25% of data is smaller than this value

Minimum value, excluding potential outliers

Potential outliers

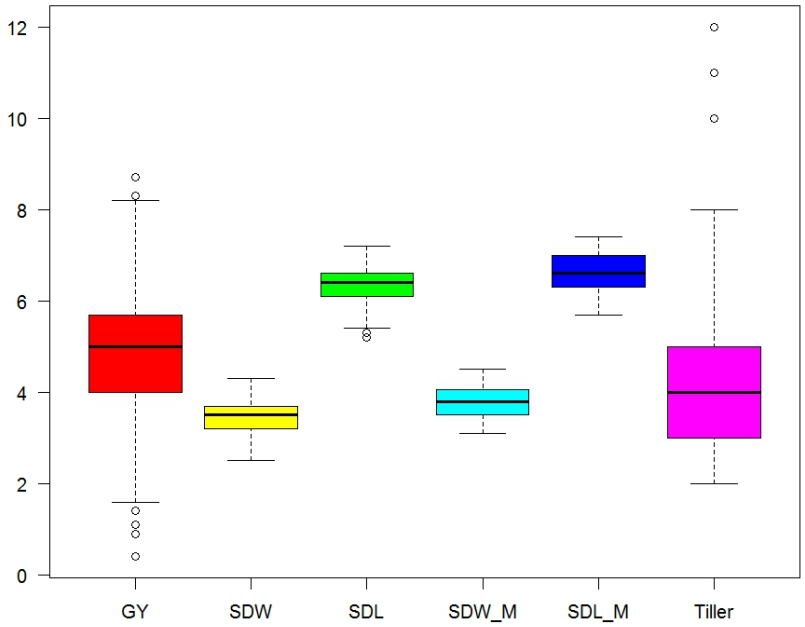
Data is not skewed distribution

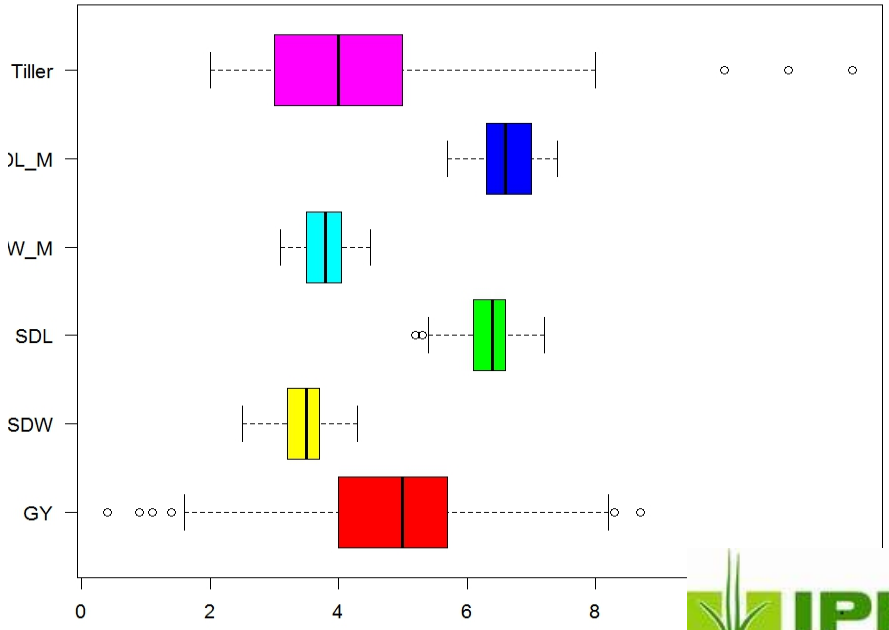
IPK GATERSLEBEN

Descriptive statistics

Five Number Summary:

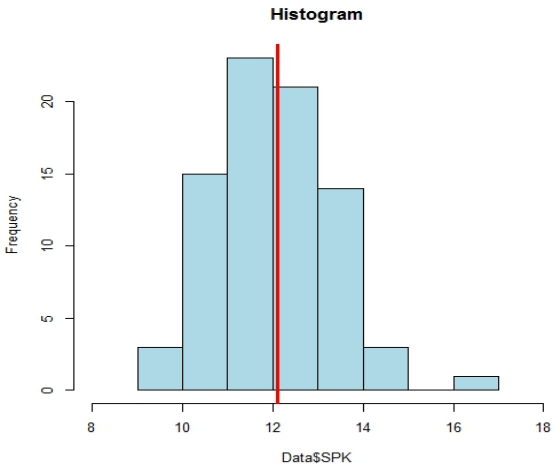
```
# box plot for table
boxplot(Data[,-c(1,3,4,7,8,9,13)],col=rainbow(7, start = 0, end = 1))
par(las=1)# all axis labels horizontal
boxplot(Data[,-c(1,3,4,7,8,9,13)],col=rainbow(7, start = 0, end = 1),
        horizontal = TRUE)
```

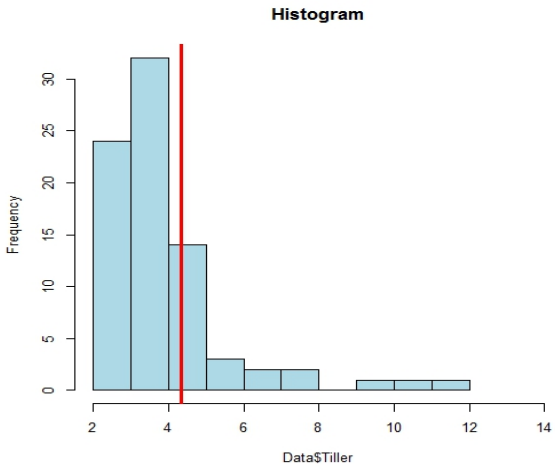


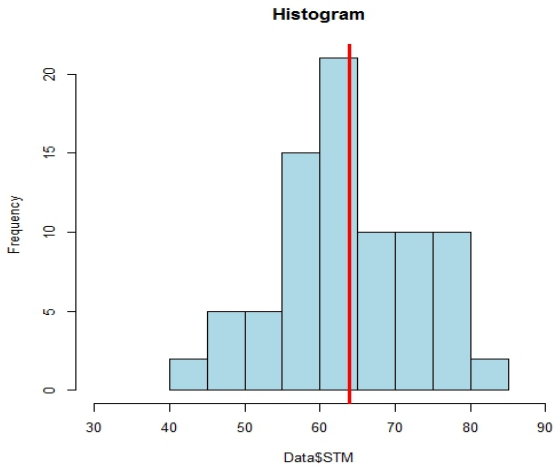


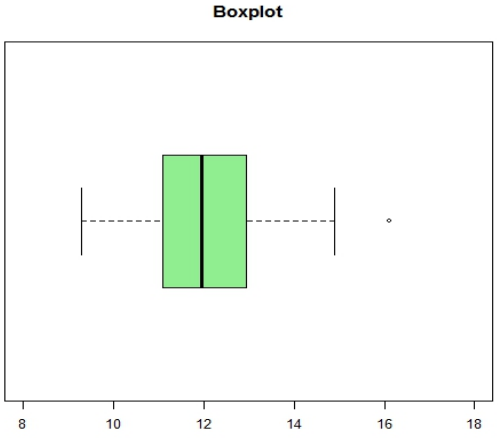
Descriptive statistics

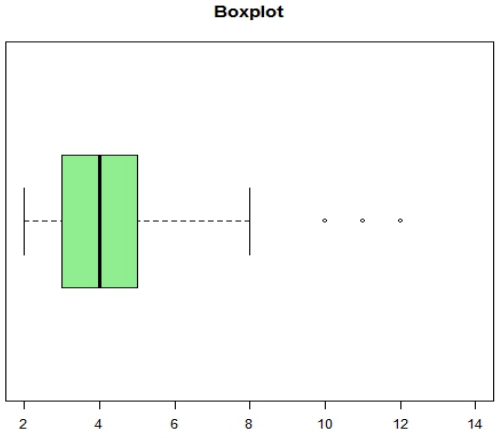
Comparing histogram and boxplot when data is **skewed distribution**:

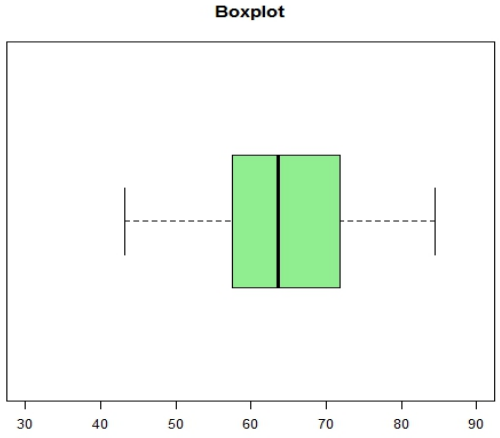












Descriptive statistics

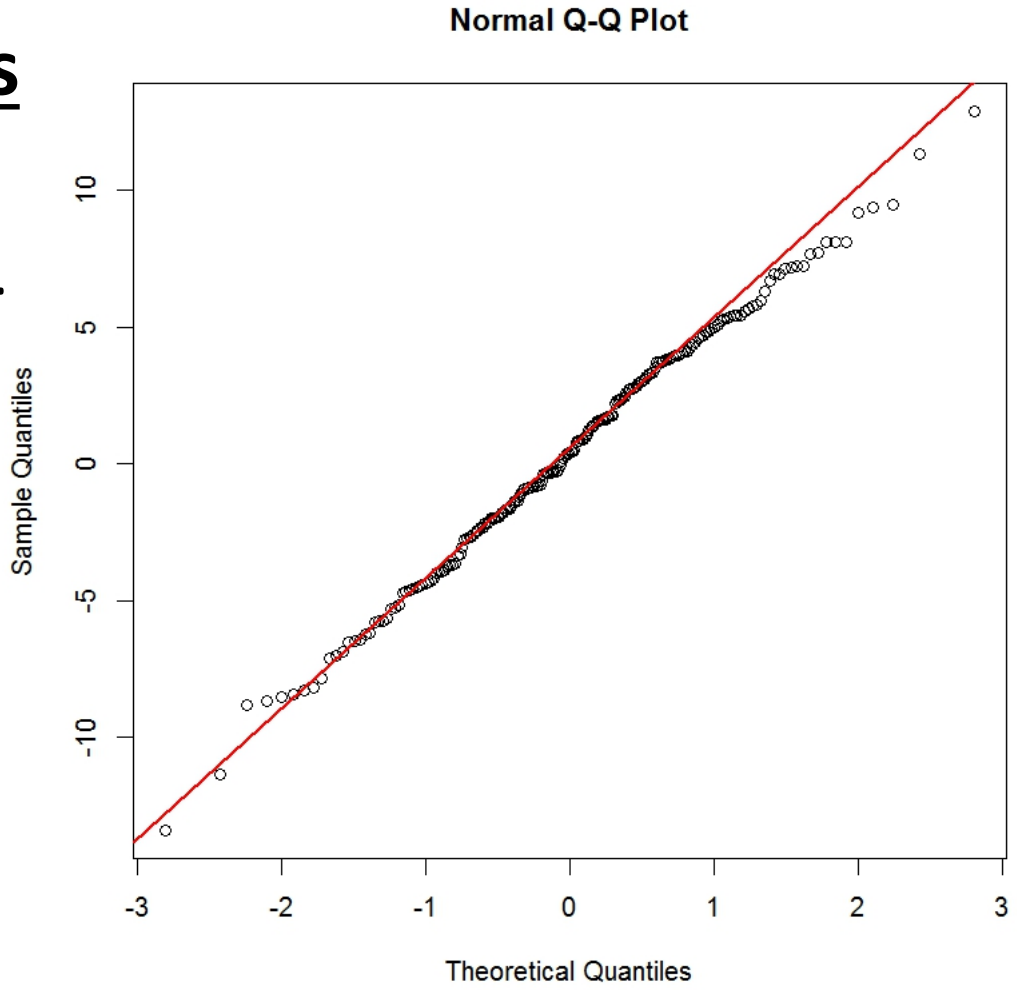
We need to check whether data is normally distributed.

Quantile-quantile plot (**QQ-plot) is used for graphically test the normality of data**

R function:

“qqnorm”: the default method produces a normal QQ-plot of the variable.

“qqline”: adds a line which passes through the first and third quartiles to a normal QQ-plot .



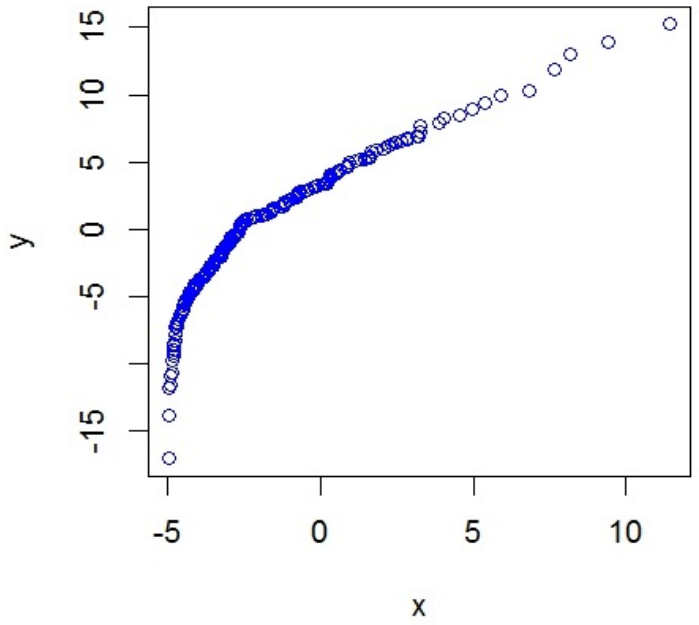

```
y <- rnorm(500,0,5)
qqnorm(y);
qqline(y, col = "red",lwd=2)
```

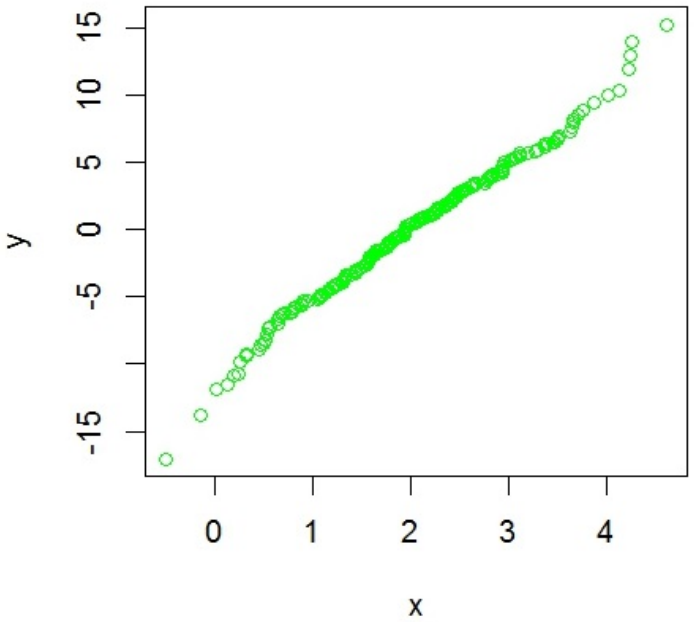
Descriptive statistics

"qqplot": QQ plot of two datasets.

Q–Q plots compare distributions, two data can be not pairwise, and the numbers of values in two data can also be not equal.

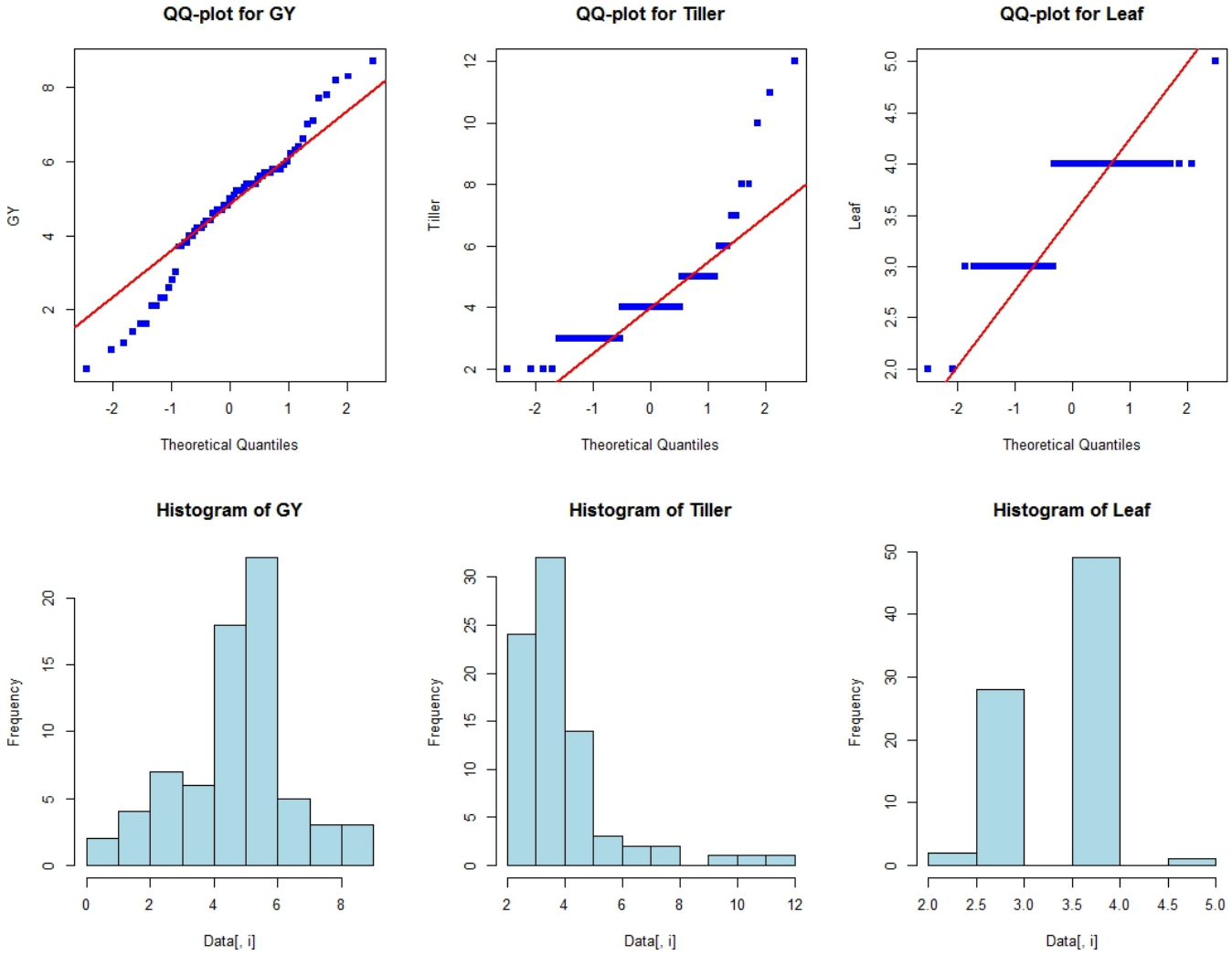
```
layout(matrix(c(1:2),2,1))
x<-rchisq(200, 2,1)-5
qqplot(x,y,col="blue")
x<-rnorm(200, 2,1)
qqplot(x,y,col="green")
```





Descriptive statistics

QQ plot and histogram for our example data



Descriptive statistics

QQ plot and histogram for our example data

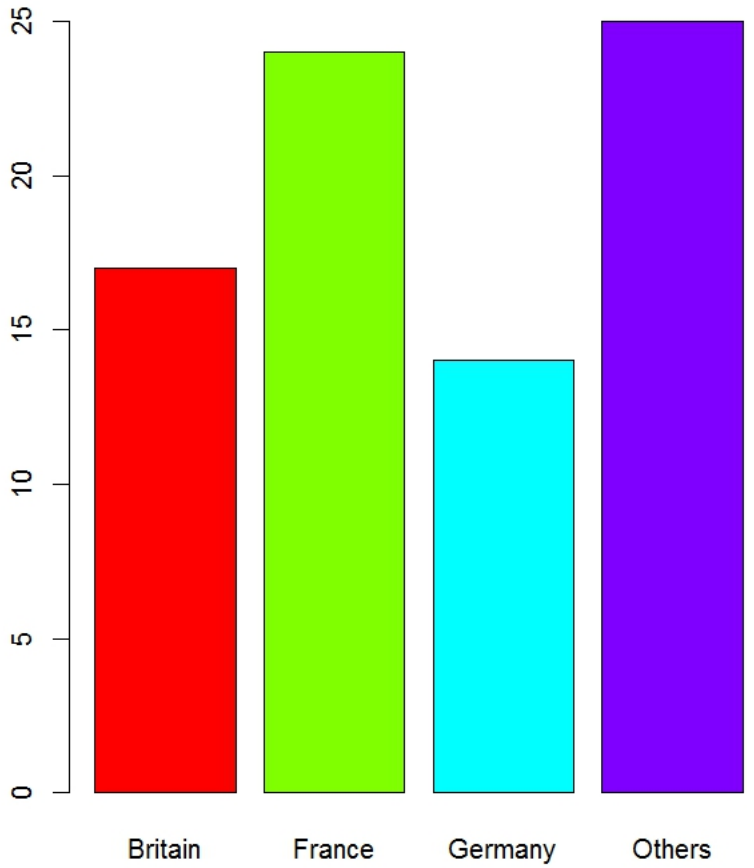
```
 # QQ-plot and histogram
layout(matrix(c(1:6),2,3))
for(i in c(2,12,13)){
qqnorm(Data[,i], ylab = paste(colnames(Data)[i]),
main=paste("QQ-plot for",colnames(Data)[i]),pch=15,col="blue")
qqline(Data[,i], col = "red",lwd=2)
hist(Data[,i],col="lightblue",main=paste("Histogram
of",colnames(Data)[i]))
}
```

Descriptive statistics

Categorical variables, barplot,

require(grDevices) # for colours

barplot(table(Data[,1]),col=rainbow(4))



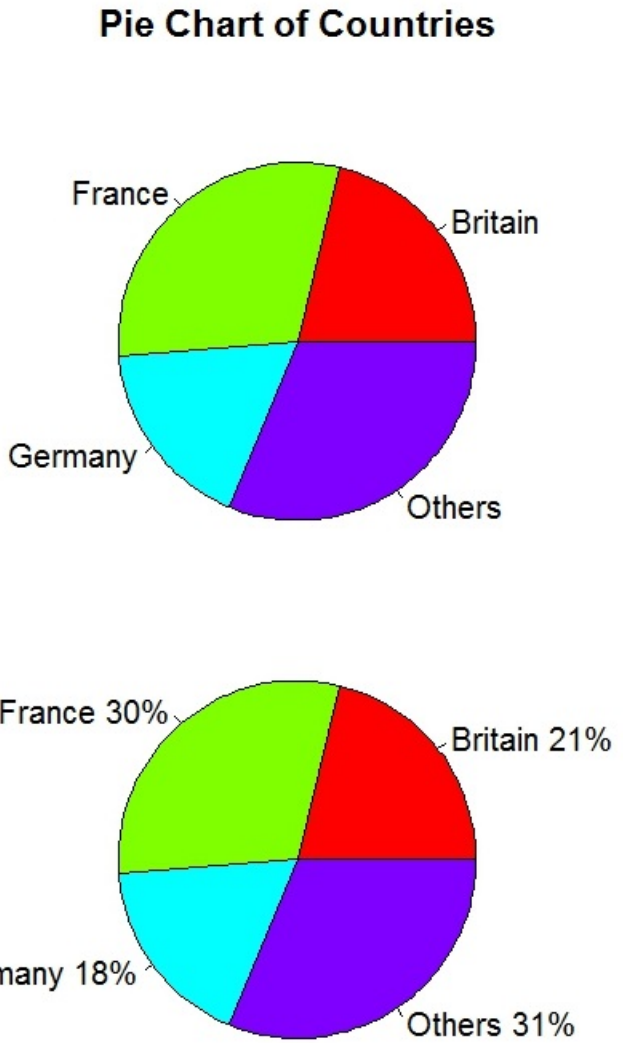
	Country	Tiller	Leaf
Gen_1	Germany	3	4
Gen_11	Germany	5	3
Gen_21	France	4	3
Gen_31	France	4	3
Gen_41	Britain	3	4
Gen_51	Britain	3	4
Gen_61	Others	4	4
Gen_71	Others	4	4

Descriptive statistics

Categorical variables, Pie Chart

```
# Simple Pie Chart
layout(matrix(c(1,2),2,1))
pie(table(Data[,1]), main="Pie Chart of
Countries",col=rainbow(4))

# add pecentage to Pie chart
Coun<- table(Data[,1])
pct <- round(Coun/sum(Coun)*100)
lb <- paste(names(table(Data[,1])),"  ",
pct,"%",sep="") # add percents to labels
pie(table(Data[,1]),labels = lb, main="Pie Chart
of Countries",col=rainbow(4))
```

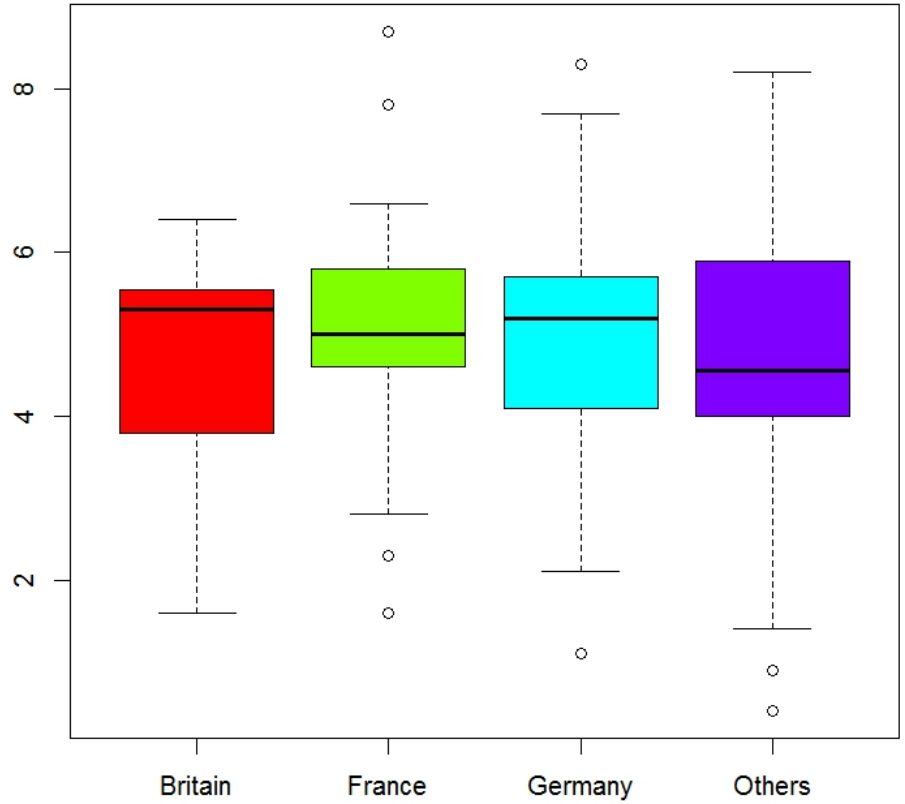


Descriptive statistics

Categorical variables, boxplot for groups

boxplot for groups

boxplot(GY ~ Country, data=Data, col=rainbow(4))

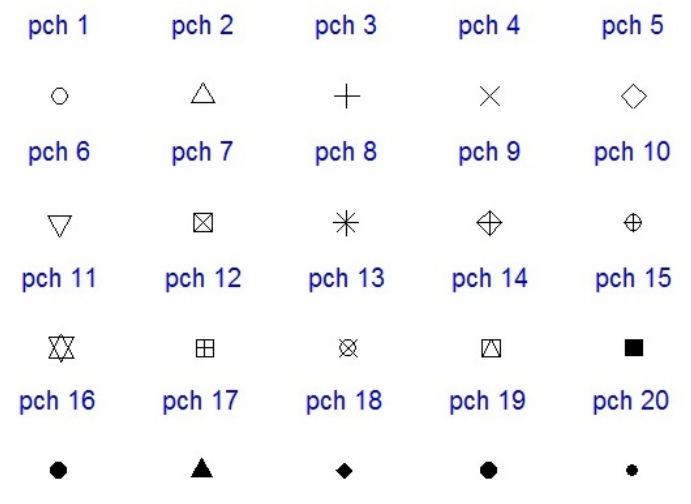


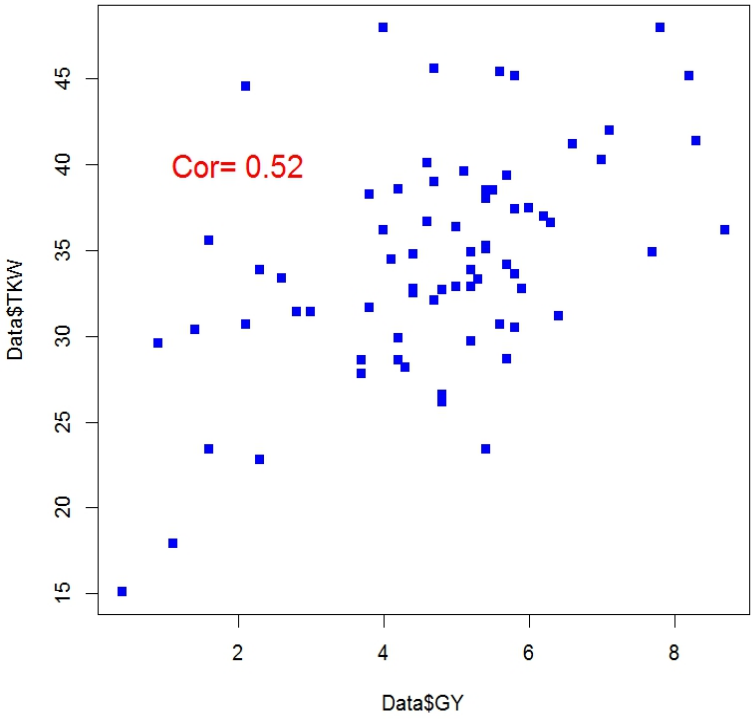
Inferential Statistics

Correlation

Correlation,

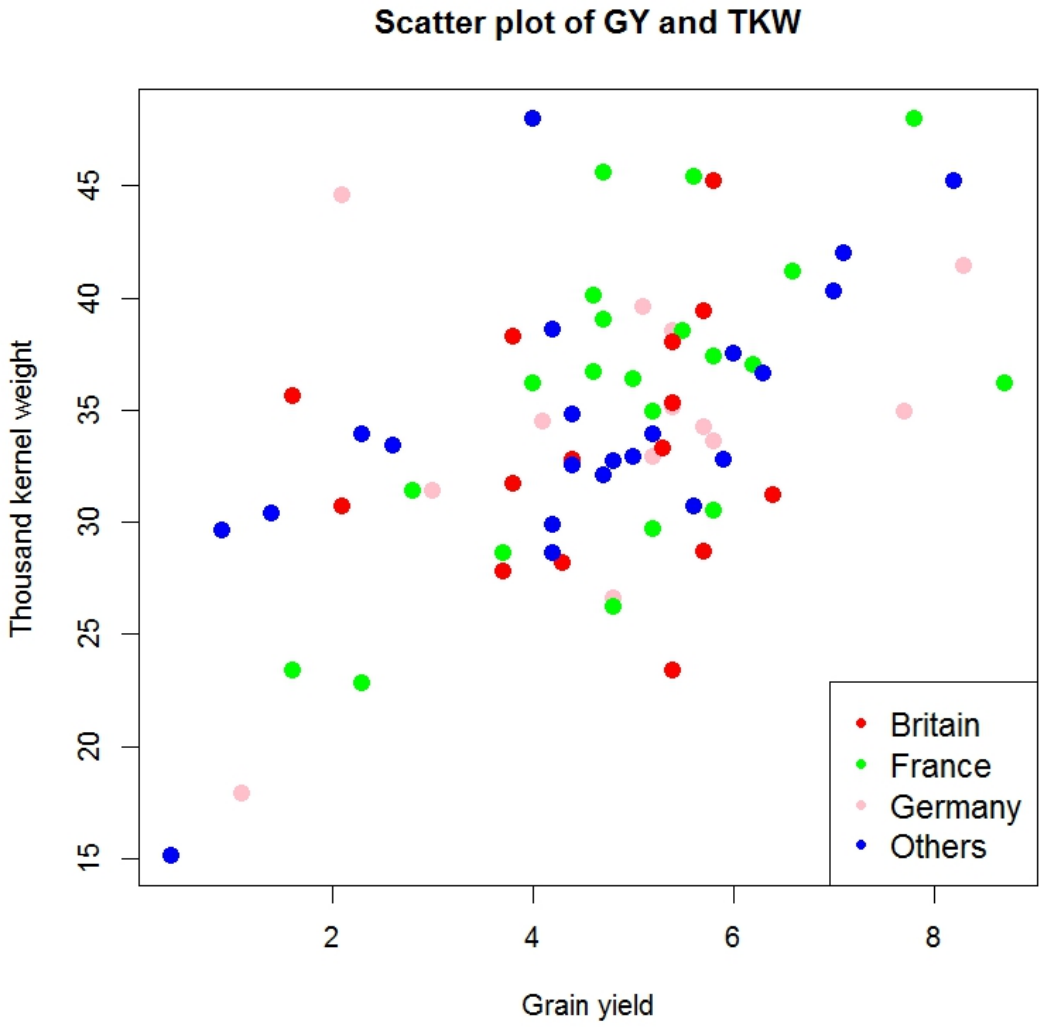
```
> cor(Data$GY,Data$TKW)
[1] NA
>cor(Data$GY,Data$TKW,"pairwise.complete.obs")
[1] 0.5201486
##simple scatter plot
plot(Data$GY,Data$TKW,col="blue",pch=15)
COR<-cor(Data$GY,Data$TKW,"pairwise.complete.obs")
text(2,40,paste("Cor=",round(COR,2)),cex=1.5,col="red")
```





Inferential Statistics

Add color to the plot according to the group



Inferential Statistics

```
# Add color to the plot according to the group
library(grDevices)
# create a function with self-defined color
table(Data$Country)
col_func <- colorRampPalette(
  colors = c("red","green","pink", "blue"),
  space = "Lab"
  )
# set country of data to define color
Own_col <- col_func(nlevels(Data$Country))
```

Inferential Statistics

```
# Add color to the plot according to different groups
plot ( x = Data$GY,y = Data$TKW, xlab = "Grain yield", ylab =
"Thousand kernel weight", cex=2, main="Scatter plot of GY and
TKW",
  pch = 20, # choose a type of dots
  col = Own_col[Data$Country] )
## add legend to data
legend( x ="bottomright",
  legend = levels(Data$Country), # for text of legend
  col = Own_col,
  pch = 20,
  cex = 1.2    # size of the legend
   )
```

X= "bottom","bottomleft","left", "topleft", "top", "topright", "right" and "center"

Inferential Statistics

Pairwise correlation

```
> cor(Data[,11:13])
      SDL_M     Tiller       Leaf
SDL_M     1         NA         NA
Tiller    NA 1.00000000 0.03103927
Leaf      NA 0.03103927 1.00000000
> cor(Data[,11:13],use="pairwise.complete.obs")
            SDL_M      Tiller       Leaf
SDL_M   1.00000000 -0.06746415 0.17638119
Tiller -0.06746415  1.00000000 0.03103927
Leaf    0.17638119  0.03103927 1.00000000
>
```

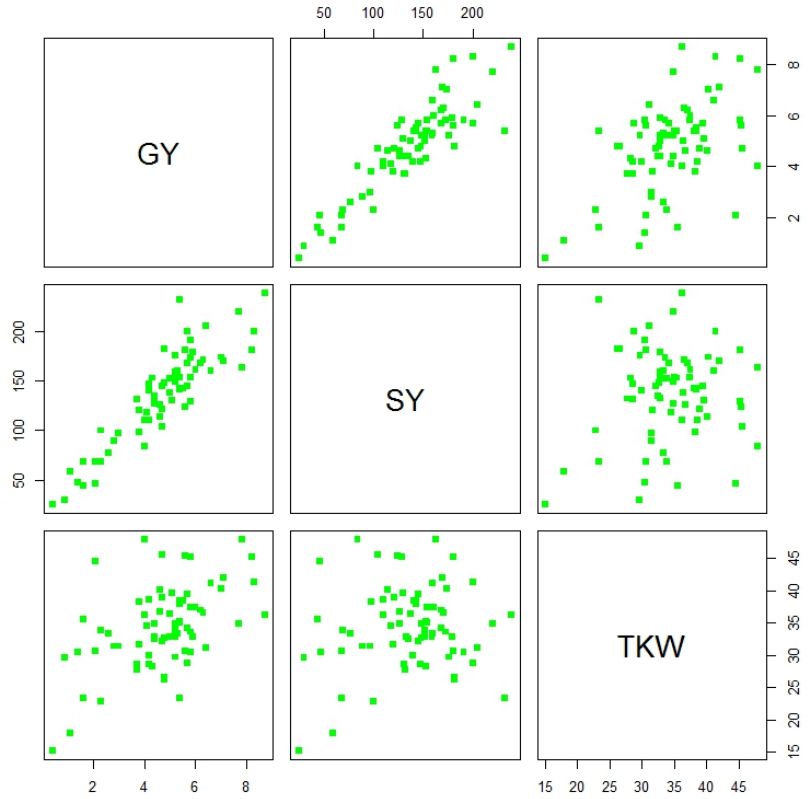
Inferential Statistics

Pairwise correlation and plot

```
> COR<- cor(Data[,2:4],use="pairwise.complete.obs")
> round(COR,2)
      GY   SY  TKW
GY  1.00 0.89 0.52
SY  0.89 1.00 0.12
TKW 0.52 0.12 1.00
```

pairwise plot

```
> pairs(Data[,2:4],pch=15,col="green")
```



Statistical testing

1. T-test (one-sample, two-sample, paired, one-tailed),

2. F-test,

3. Correlation test,

4. Chi square test,

5. Fisher exact test

Statistical testing

1. T-test (one-sample, two-sample, paired),

	Country	GY	SY	TKW	SDW	SDL	STM	SPK	TKW_M	SDW_M	SDL_M	Tiller	Leaf
Gen_1	Germany	5.1	130	39.6	3.5	6	62.1	11.3	45.3	4.2	6.4	3	4
Gen_2	Germany	5.7	168	34.2	3.6	6.6	68.7	12.4	44	4.2	6.3	11	3
Gen_3	Germany	4.8	182	26.6	3	6.5	47.4	14	32.8	3.8	6.5	4	3
Gen_4	…	…	…	…	…	…	…	…	…	…	…	…	…

T-test is usually used to compare the means of two samples

Question 1(one sample): Is the average leaf number per stem in this data significantly different from 4?

Question 2(two sample): Is there significant difference between grain yield of genotypes from Germany and from France?

Question 3(paired): Is there significant difference between Seed width(SDW) and Seed width of main spike(SDW_M)?

Question 4(paired): Is the Seed width(SDW) significantly smaller or larger than Seed width of main spike(SDW_M)?

IPK GATERSLEBEN

Statistical testing

One-sample t-test,

Question 1(one sample): Is the average leaf number per stem in this data significantly different from 4?

R function: t.test()

```
> t.test(Data$Leaf,mu=4)
        One Sample t-test
data:  Data$Leaf
t = -6.1627, df = 79, p-value = 2.829e-08
alternative hypothesis: true mean is not equal to 4
95 percent confidence interval:
 3.487343 3.737657
sample estimates:
mean of x
   3.6125
```

p-value = 2.829e-08=$2.829*10^{-8}$
P value<0.001, so it is significantly different from 4

The mean of data is only from one sample, and the "true mean" should locate in this range with 95% probability.

Statistical testing

Two-sample t-test

Question 2(two sample): Is there significant difference between grain yield of the genotypes from Germany and France?

```
> Y1<-Data$GY[which(Data$Country=="Germany")]
> Y2<-Data$GY[which(Data$Country=="France")]
> t.test(Y1,Y2)
        Welch Two Sample t-test
data:  Y1 and Y2
t = -0.16561, df = 21.9, p-value = 0.87
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.481448  1.262401
sample estimates:
mean of x mean of y
 4.900000  5.009524
```

p-value =0.87

P value>0.05, **no significant difference**

Statistical testing

Paired t-test

Question 3(paired): Is there significant difference between Seed width(SDW) and Seed width of main spike(SDW_M)?

```
> mean(Data$SDW,na.rm=TRUE);mean(Data$SDW_M,na.rm=TRUE)
[1] 3.484932 ; [1] 3.790278
> t.test(Data$SDW,Data$SDW_M,paired=TRUE)

        Paired t-test

data:  Data$SDW and Data$SDW_M
t = -6.3664, df = 71, p-value = 1.667e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3994308 -0.2089025
sample estimates:
mean of the differences
             -0.3041667
```

p-value = 1.667e-08=$1.667*10^{-8}$
P value<0.001, **significant difference**

Statistical testing

One tailed t-test

Question 4(paired): Is the Seed width(SDW) significantly smaller than Seed width of main spike(SDW_M)?

```
> t.test(Data$SDW,Data$SDW_M,paired=TRUE, alt="less")
Paired t-test
data:  Data$SDW and Data$SDW_M
t = -6.3664, df = 71, p-value = 8.333e-09
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -0.2245419
sample estimates:
mean of the differences
           -0.3041667
```

p-value = 1.667e-08=1.667*10^{-8}

P value<0.001, **significantly smaller**

Statistical testing

One tailed t-test

Question 4(paired): Is the Seed width(SDW) significantly larger than Seed width of main spike(SDW_M)?

```
> t.test(Data$SDW,Data$SDW_M,paired=TRUE, alt="greater")

Paired t-test

data:  Data$SDW and Data$SDW_M
t = -6.3664, df = 71, p-value = 1
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.3837914       Inf
sample estimates:
mean of the differences
             -0.3041667

>
```

p-value =1
Not significantly larger

Statistical testing

2. F-test

F-test is usually used to compare the variances of two variables

Question 1: Is there significant difference between variance of grain yield of the genotypes from Germany and France?

Question 2: Is the variance of Seed width(SDW) significantly smaller or larger than Seed width of main spike(SDW_M)?

Statistical testing

F-test(two tailed)

Question 1: Is there significant difference between variance of grain yield of the genotypes from Germany and France?

```
> Y1<-Data$GY[which(Data$Country=="Germany")]
> Y2<-Data$GY[which(Data$Country=="France")]
> var.test(Y1,Y2)
    F test to compare two variances
data:  Y1 and Y2
F = 1.4707, num df = 12, denom df = 20, p-value = 0.4306
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5496281 4.5191627
sample estimates:
ratio of variances
         1.470712
```

```
> var(Y1,na.rm=TRUE);var(Y2,na.rm=TRUE)
[1] 4.001667; [1] 2.720905
```

p-value >0.05
No significant difference

Statistical testing

F-test(one tailed)

Question 2: Is the variance of Seed width(SDW) significantly smaller or larger than Seed width of main spike(SDW_M)?

```
> var.test(Data$SDW,Data$SDW_M, alternative = "less")#, "greater","two.sided"
      F test to compare two variances
data:  Data$SDW and Data$SDW_M
F = 1.1214, num df = 72, denom df = 71, p-value = 0.685
alternative hypothesis: true ratio of variances is less than 1
95 percent confidence interval:
 0.000000 1.658459
sample estimates:
ratio of variances
          1.121359
```

var(Data$SDW,na.rm=TRUE);
var(Data$SDW_M,na.rm=TRUE)
[1] 0.123242 ; [1] 0.1099041

p-value >0.05
Not significantly lower

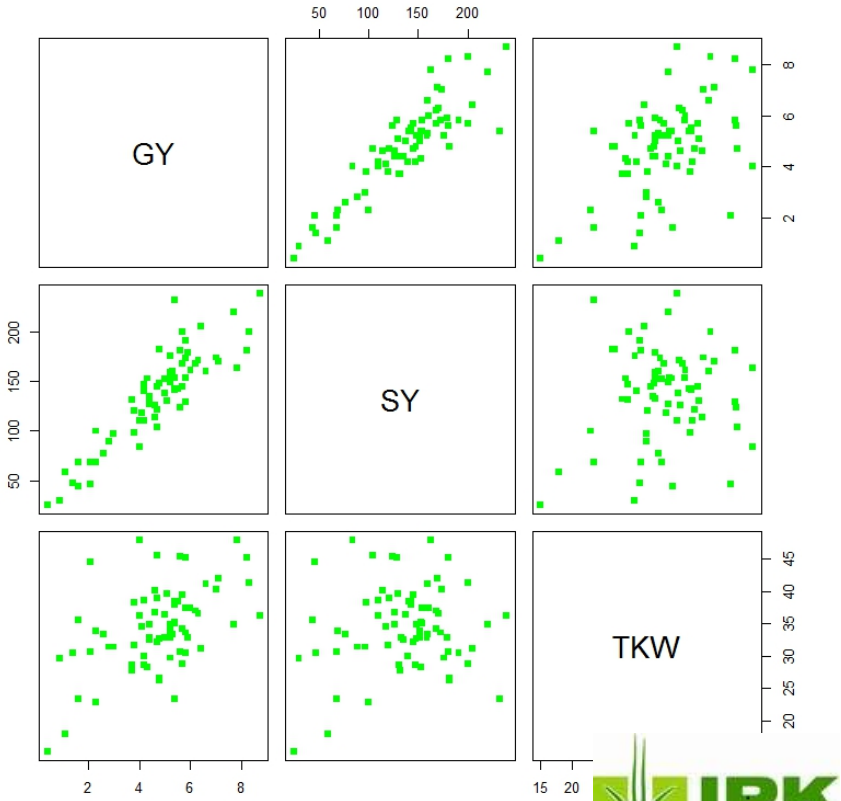
Statistical testing

3. Correlation test

Correlation test is usually used to estimate whether there is linear association between two variables

Question 1: Is there a **linear association** between grain yield and TKW?

Question 2: What is the effect of sample size on the correlation test above?



Statistical testing

Correlation test

Question 1: Is there a linear association between grain yield and TKW?

```
> cor.test(Data$GY,Data$TKW)

   Pearson's product-moment correlation

data:  Data$GY and Data$TKW
t = 5.0589, df = 69, p-value = 3.333e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3264618 0.6719138
sample estimates:
      cor
0.5201486
```

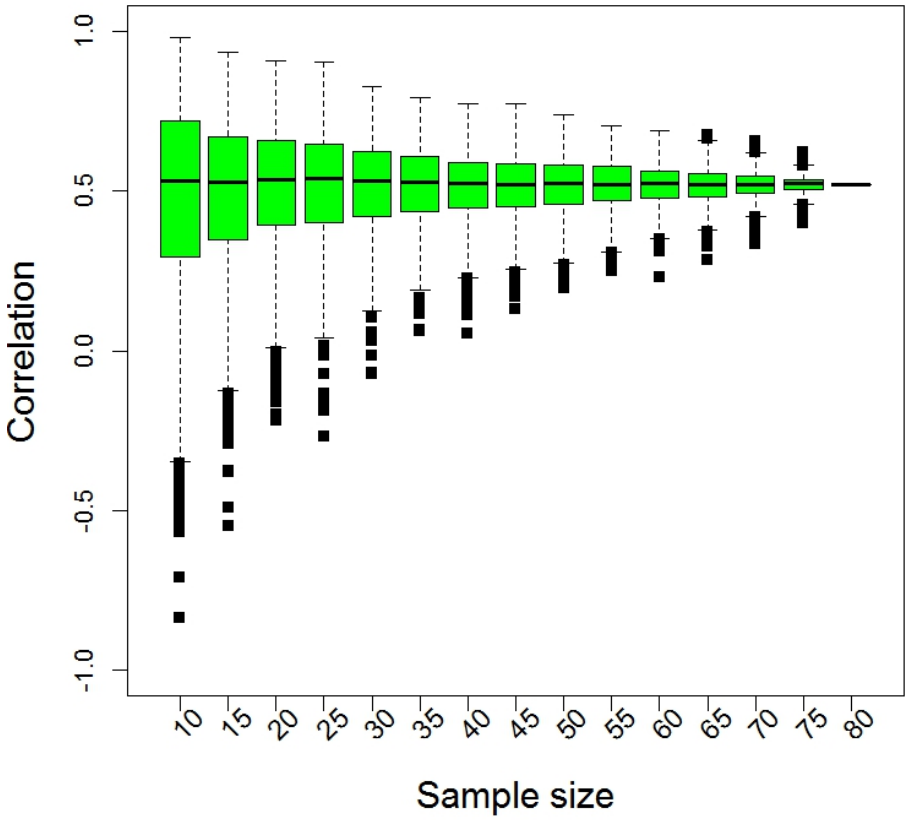
p-value <0.05
significantly correlated

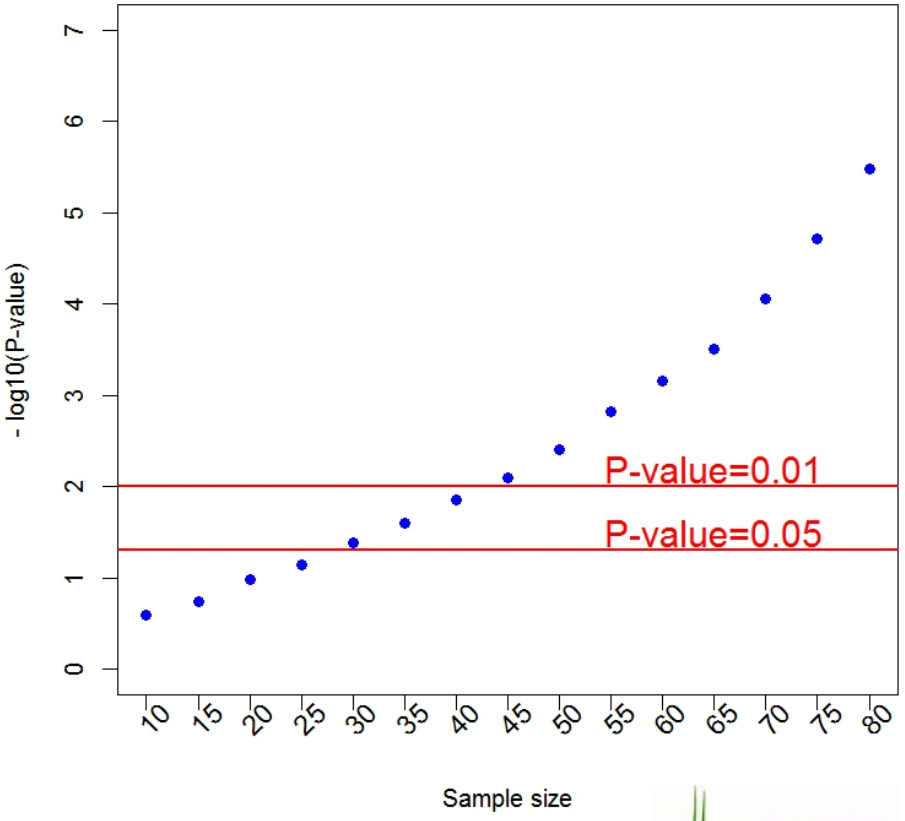
The Pearson's correlation coefficient between GY and TKW is -0.52 and the p-value is $3.333 * 10^{-6}$

Statistical testing

Correlation test

Question 2: What is the effect of sample size on the correlation test above?(simulate 1000 times)





Statistical testing

Correlation test

Question 2: What is the effect of sample size on the correlation test above?

```
> #Power analysis for correlation
> pwr.r.test(n = NULL, r = 0.52, sig.level = 0.05,  power = 0.80, alternative = "two.sided")

     approximate correlation power calculation (arctangh transformation)

              n = 25.87738
              r = 0.52
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

If we change sig.level to 0.01, then we will get n = 37.746

Statistical testing

4. Chi square test

Chi-Square test for “Goodness-of-Fit “

Question 1: Supposing leaf number (L) of whole wheat population has frequencies for L≤3 and L>3 are 25% and 75%, respectively. Then how 'good' can our data 'fit' the assumed frequencies?

	Leaf number(L)		
	L≤3	**L>3**	**All**
Our	30	50	80
	37.5%	62.5%	
Theory	25%	75%	

Example data

Chi-Square test for "Goodness-of-Fit "

```
> ## Chi square test for "Goodness-of-Fit "
> observed = c(30, 50)        # observed frequencies
> expected = c(0.25, 0.75)    # expected proportions
> chisq.test(x = observed,p = expected)

        Chi-squared test for given probabilities

data:  observed
X-squared = 6.6667, df = 1, p-value = 0.009823
```

	Leaf number(L)		
	L≤3	**L>3**	**All**
Our	30	50	80
Theory	25%	75%	

Statistical testing

Two random variables x and y are **independent** if the probability distribution of one variable is not affected by the presence of the other one.

4. Chi square test

Chi-Square Test of Independence of two categorical variables

Question 2: Is tiller number of our wheat genotypes independent of their country of origin?

	Tiller number(N)		
Country	**N≤4**	**N>4**	**All**
Britain	12	5	17
France	17	7	24
Germany	8	6	14
Others	19	6	25

IPK GATERSLEBEN

Statistical testing

Chi-Square Test of Independence of two categorical variables

```
> tbl<-tabular(Country~Factor(Tiller>4,"Leaf number")+1,data=Data)
> Tiller<-
data.frame(matrix(unlist(tbl),4,3),row.names=names(table(Data$Country)))
> chisq.test(Tiller[,1:2])
        Pearson's Chi-squared test
data:  Tiller[, 1:2]
X-squared = 1.5413, df = 3, p-value = 0.6728
Warning message:
In chisq.test(Tiller[, 1:2]) : Chi-squared approximation may be incorrect
```

P-value > 0.05 significance level, we do not reject the null hypothesis that the tiller number is independent of the origin country.

The warning message is due to too much small cell values (should large than 5) in the contingency table. To avoid such warning, we have to use Fisher exact test instead of Chi square test.

Statistical testing

Chi square test of Independence

```
>Tiller<-data.frame(matrix(c(37,19,18,6),2,2),
  row.names=c("Europe","Others"))
> chisq.test(Tiller)
        Pearson's Chi-squared test with Yates' continuity correction
data:  Tiller
X-squared = 0.27706, df = 1, p-value = 0.5986
```

	Tiller number(N)		
	N≤4	**N>4**	**All**
Britain, France and Germany	37	18	55
Others	19	6	25

Statistical testing

4. Chi square test

An other example

	Tiller number(N)		
Country	**N≤4**	**N>4**	**All**
Britain	12	5	17
France	17	7	24
Germany	8	16	24
Others	19	6	25

```
> Tiller[3,2]<-16
> chisq.test(Tiller[,1:2])
        Pearson's Chi-squared test
data:  Tiller[, 1:2]
X-squared = 11.803, df = 3, p-value = 0.008089
>
```

Statistical testing

5. Fisher exact test

Fisher exact test of Independence of two categorical variables

Question: Is tiller number of our wheat genotypes independent of their country of origin?

	Tiller number(N)		
Country	**N≤4**	**N>4**	**All**
Britain	12	5	17
France	17	7	24
Germany	8	6	14
Others	19	6	25

Statistical testing

5. Fisher exact test

```
> # fisher exact test
> tbl<-tabular(Country~Factor(Tiller>4,"Leaf number")+1,data=Data)
> Tiller<-data.frame(matrix(unlist(tbl),4,3),row.names=names(table(Data$Country)))
> fisher.test(Tiller[,1:2],alternative="two.sided")

        Fisher's Exact Test for Count Data

data:  Tiller[, 1:2]
p-value = 0.6952
alternative hypothesis: two.sided
```

Linear regression

- **What is linear regression?**
  - The regression equation:
  - Y = a + bX + e
  - where Y is the dependent or criterion variable, a is the intercept or point, X is the independent or predictor variable, a and b are often called “regression coefficients.
- **Assumption:**
  - The X and Y are linearly related
  - $\widehat{Y}$**=a+bX is the estimated value**

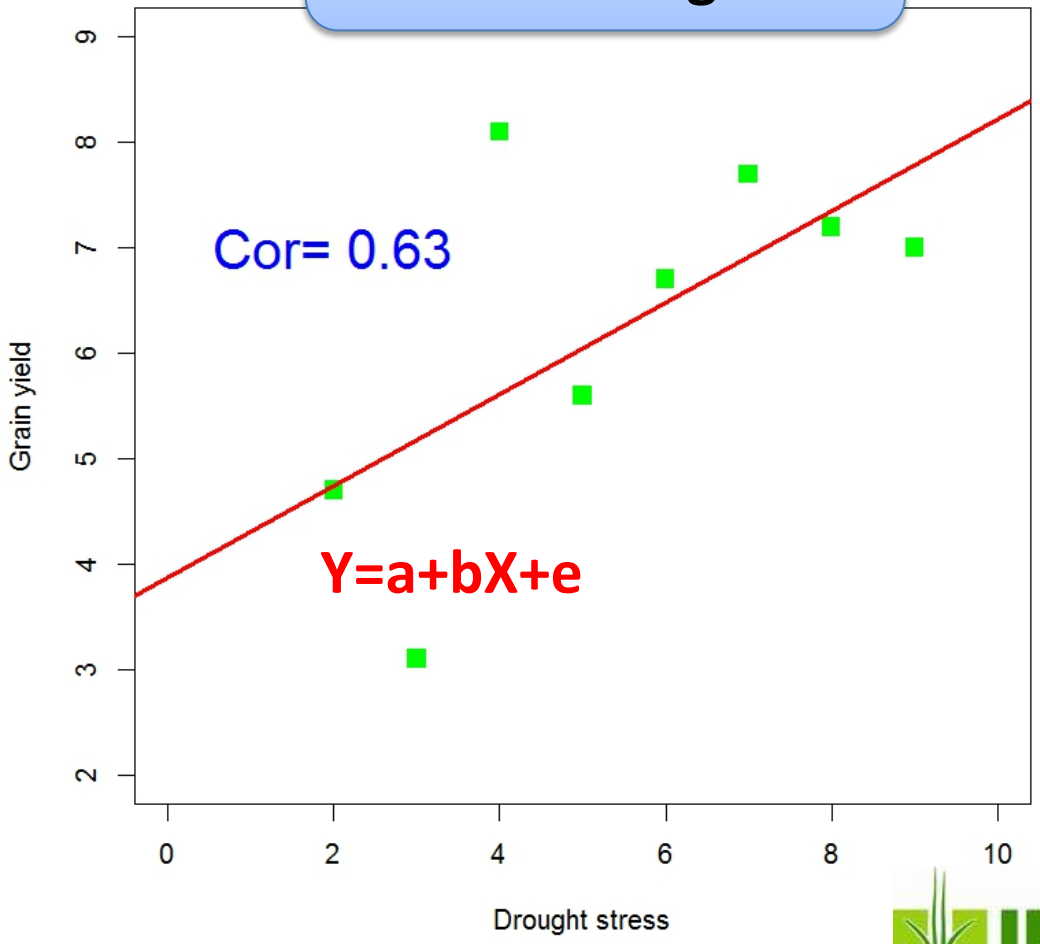
- **What is linear regression?**
  - The regression equation:
  - Y = a + bX + e
  - where Y is the dependent or criterion variable, a is the intercept or point, X is the independent or predictor variable, a and b are often called “regression coefficients.
- **Assumption:**
  - The X and Y are related linearly
  - $\widehat{Y}$=a+bX is the estimated value

A Example for linear regression

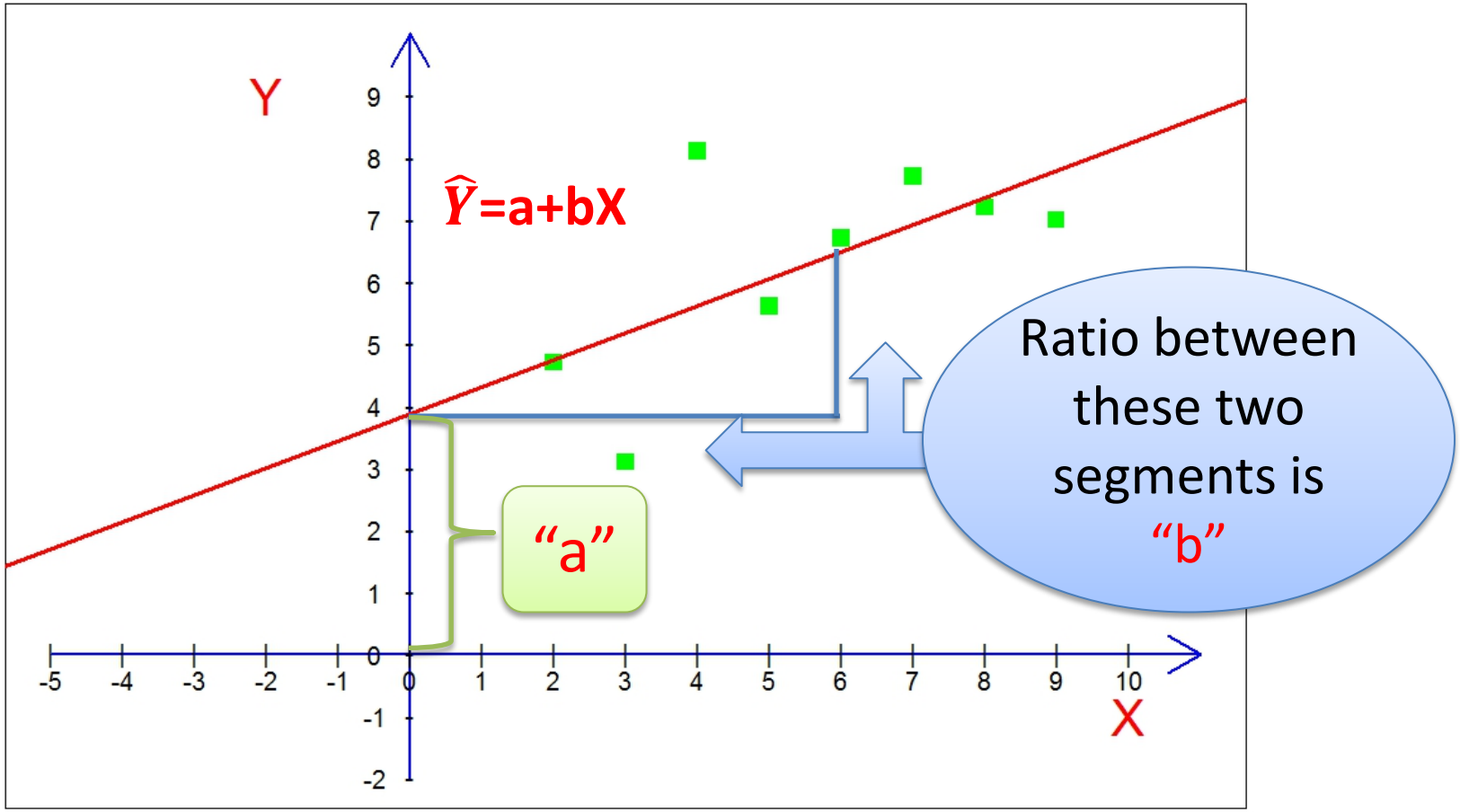
Y: Grain yield, Ton/ha
X: Drought stress, 1-9(high to lower)

Y	X
5.6	5
3.1	3
4.7	2
7.7	7
8.1	4
7	9
7.2	8
6.7	6

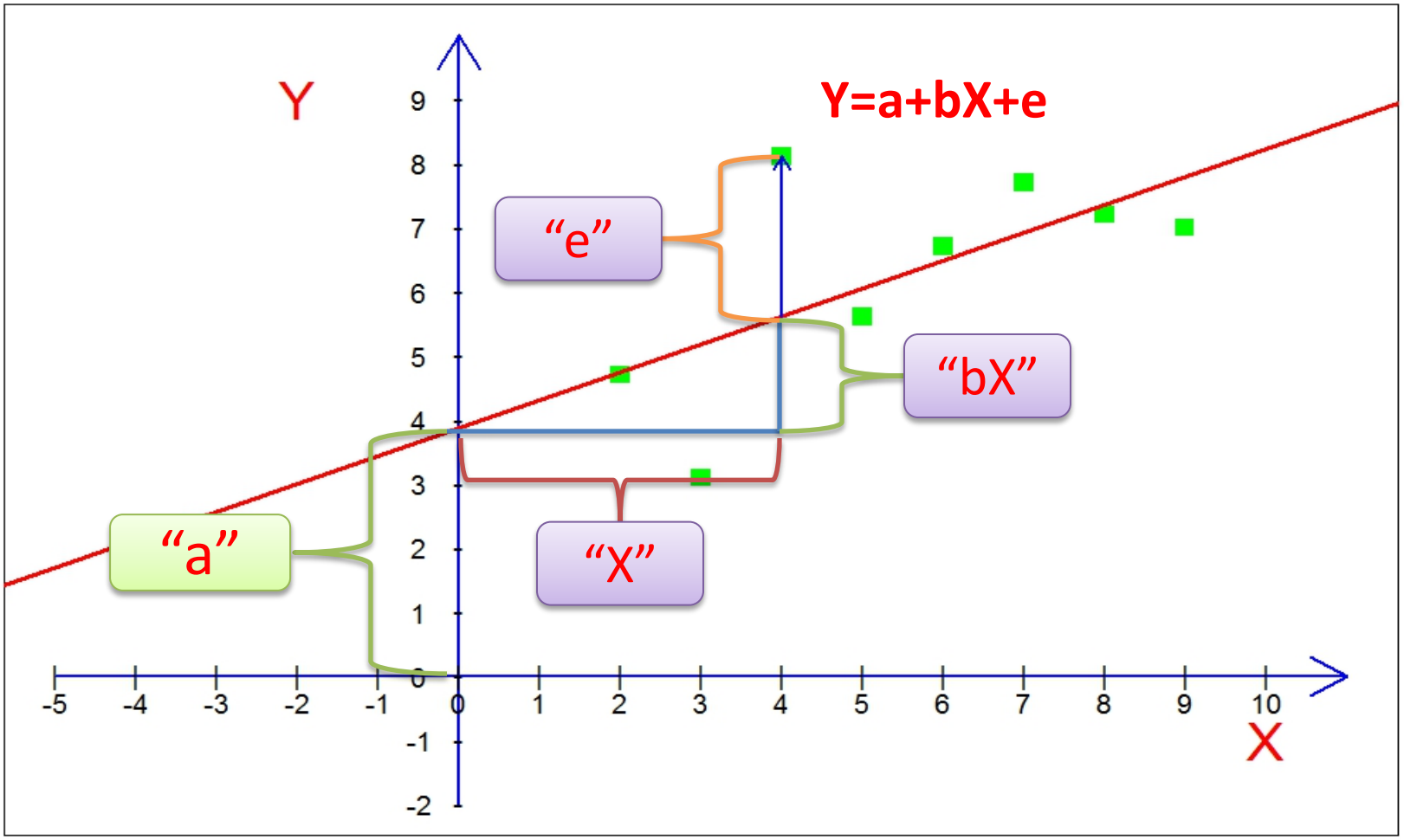
What is "a","b" and "e" in this figure?



A Example for linear regression

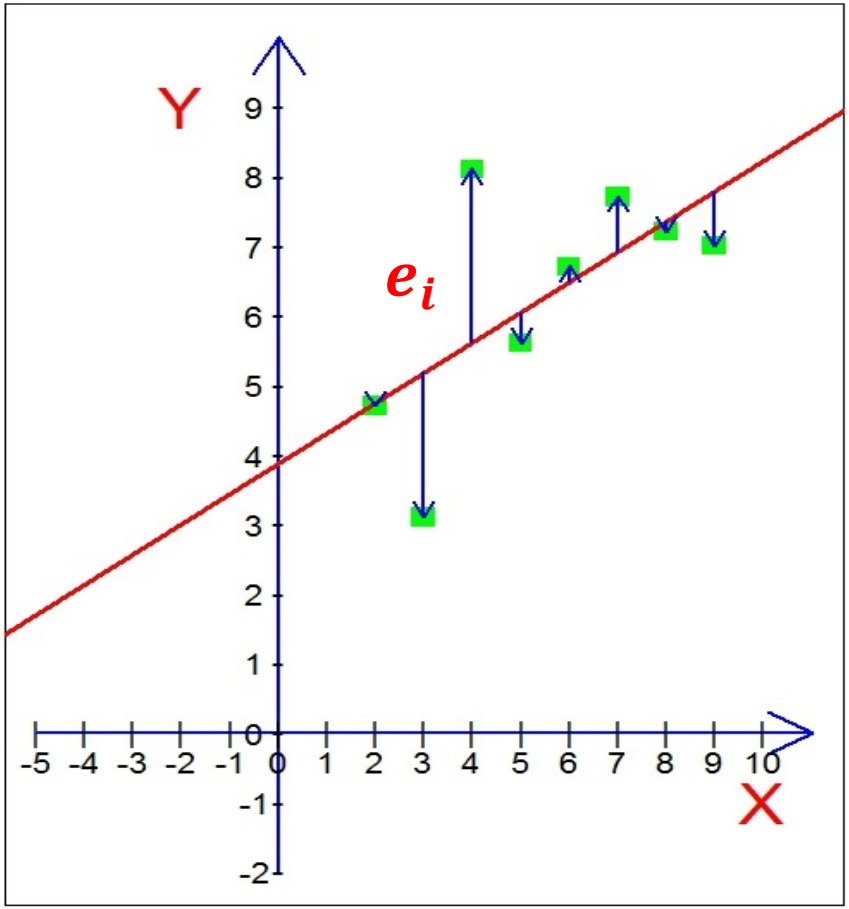


A Example for linear regression

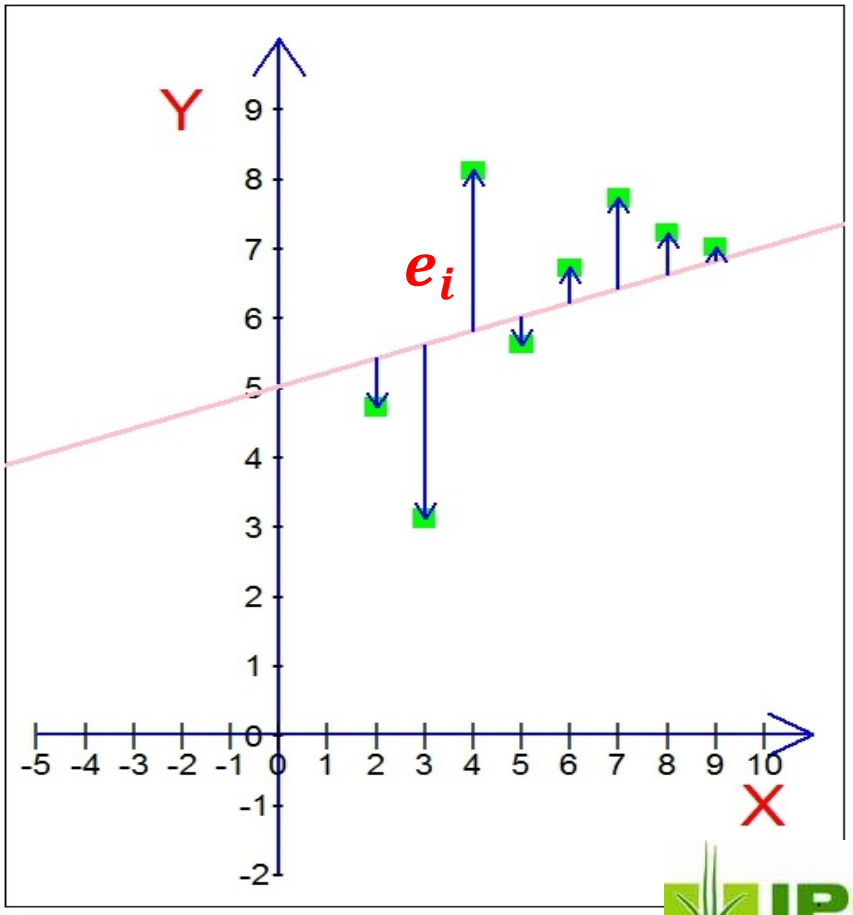


least squares solution

To minimize the error sum of squares or *residuals:* $\sum_{i=1}^{8} {e_i}^2$



Drought stress

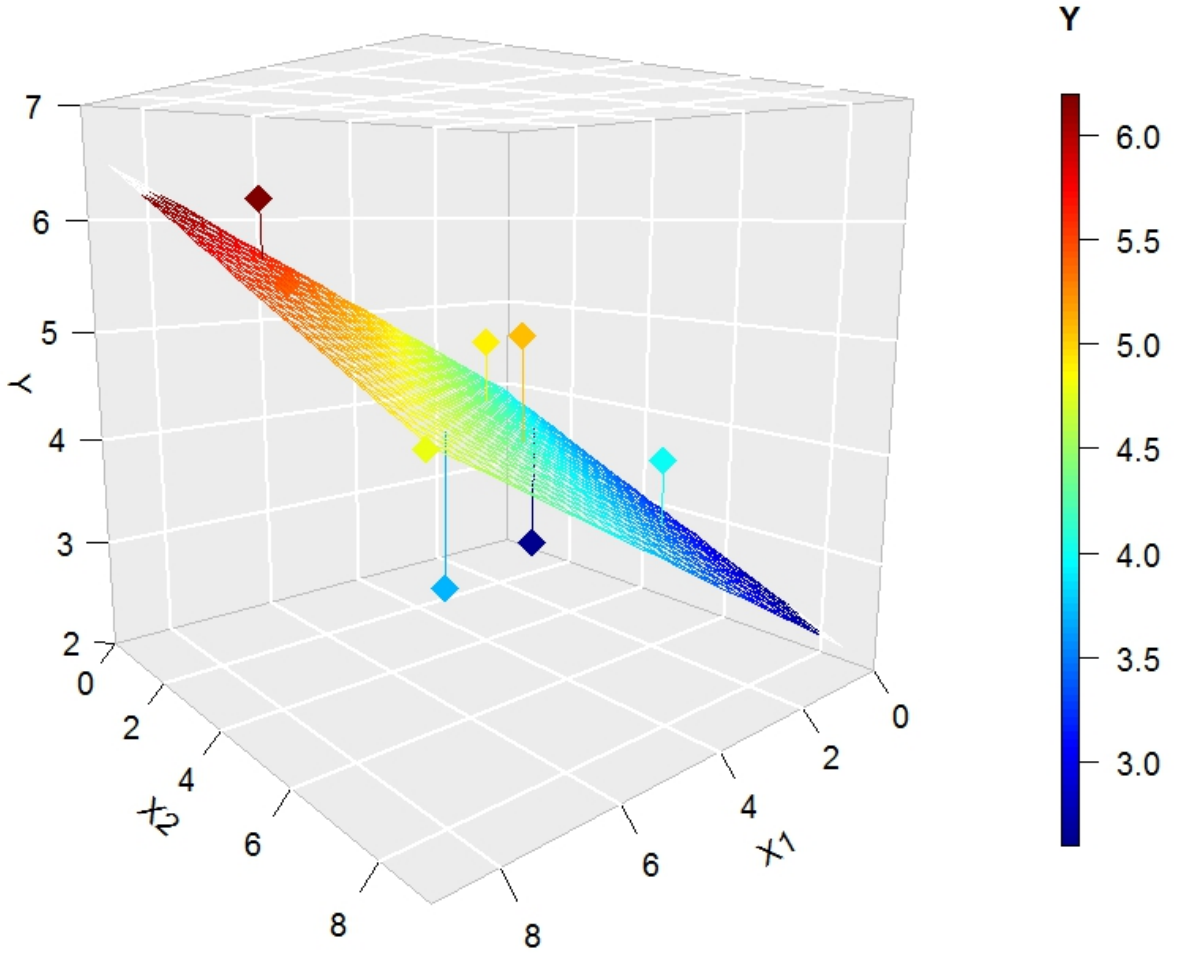


Drought stress

IPK GATERSLEBEN

least squares solution

Y	X1	X2
5.1	5	6
4	3	7
2.6	2	3
6.2	7	2
4.9	4	4
4.8	9	9
3.7	8	8
5.4	6	1



Example code

```
# Example code for linear regression, one variable
X<-c(5,3,2,7,4,9,8,6)
Y<-c(5.6,3.1,4.7,7.7,8.1,7.0,7.2,6.7)
> lm(Y~X)
Call:
lm(formula = Y ~ X)
Coefficients:
(Intercept)          X
     3.8726       0.4345
```

Y	X
5.6	5
3.1	3
4.7	2
7.7	7
8.1	4
7	9
7.2	8
6.7	6

IPK GATERSLEBEN

Example code (continue)

```
> # Example code for linear regression, two variable
> X1<-c(5,3,2,7,4,9,8,6)
> X2<-c(6,7,3,2,4,9,8,1)
> Y<-c(5.1,4.0,2.6,6.2,4.9,4.8,3.7,5.4)
> lm(Y~X1+X2)

Call:
lm(formula = Y ~ X1 + X2)

Coefficients:
(Intercept)          X1          X2
     3.9472      0.3038     -0.2061
```

Y	X1	X2
5.1	5	6
4	3	7
2.6	2	3
6.2	7	2
4.9	4	4
4.8	9	9
3.7	8	8
5.4	6	1

IPK GATERSLEBEN

Example code (continue)

Question: we want to know effects of other traits on grain yield

```
# Example code for linear regression, multiple variables
> lm(Data[,-1])
Call:
lm(formula = Data[, -1])
Coefficients:
(Intercept)        SY       TKW       SDW       SDL       STM       SPK
 -3.6565105 0.0312810 0.1147524 -0.1790541 -0.1444958 0.0008549 -0.0343090
     TKW_M     SDW_M     SDL_M    Tiller      Leaf
 0.0200975 0.0427532 0.1614365 0.0305227 -0.0372396
```

	Country	GY	SY	TKW	SDW	SDL	STM	SPK	TKW_M	SDW_M	SDL_M	Tiller	Leaf
Gen_1	Germany	5.1	130	39.6	3.5	6	62.1	11.3	45.3	4.2	6.4	3	4
Gen_2	Germany	5.7	168	34.2	3.6	6.6	68.7	12.4	44	4.2	6.3	11	3
Gen_3	…	…	…	…	…	…	…	…	…	…	…	…	…

Example code (continue)

```
> # alternative code for linear regression
> R_lm<-lm(GY~SY+TKW+SDW+SDL+STM+SPK+TKW_M+SDW_M+SDL_M+Tiller+Leaf,Data)
> summary(R_lm)
Call:
lm(formula = GY ~ SY + TKW + SDW + SDL + STM + SPK + TKW_M +
    SDW_M + SDL_M + Tiller + Leaf, data = Data)
Residuals:
     Min       1Q   Median       3Q      Max
-1.07077 -0.16749 -0.04219  0.11035  1.15401
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.6565105  1.0755322  -3.400  0.00123 **
SY           0.0312810  0.0010303  30.361  < 2e-16 ***
TKW          0.1147524  0.0110670  10.369 7.97e-15 ***
SDW         -0.1790541  0.1562752  -1.146  0.25660
SDL         -0.1444958  0.1085111  -1.332  0.18819
```

Effect of seed yield

Standard error of effects

Pvalues of effect, P<0.05 means it is significantly different from 0

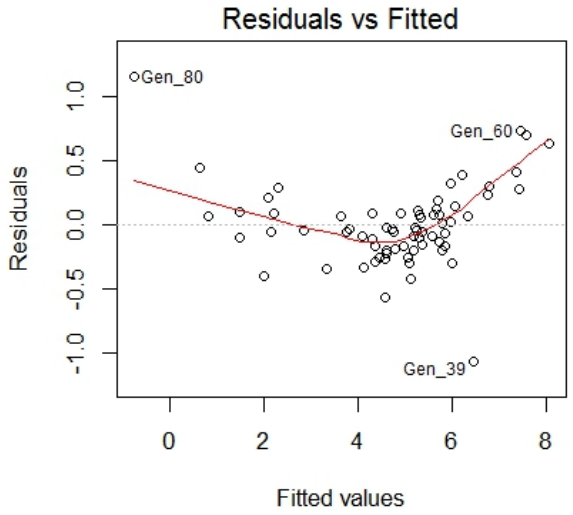
Example code (continue)

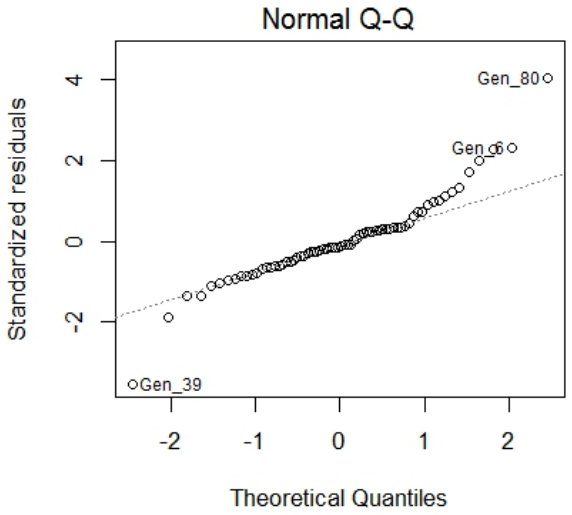
```
STM         0.0008549  0.0055743   0.153  0.87864
SPK        -0.0343090  0.0357211  -0.960  0.34081
TKW_M       0.0200975  0.0149563   1.344  0.18427
SDW_M       0.0427532  0.1511253   0.283  0.77826
SDL_M       0.1614365  0.1570665   1.028  0.30830
Tiller      0.0305227  0.0284358   1.073  0.28754
Leaf       -0.0372396  0.0859893  -0.433  0.66657
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3434 on 58 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.9675,    Adjusted R-squared:  0.9613
F-statistic:   157 on 11 and 58 DF,  p-value: < 2.2e-16
```

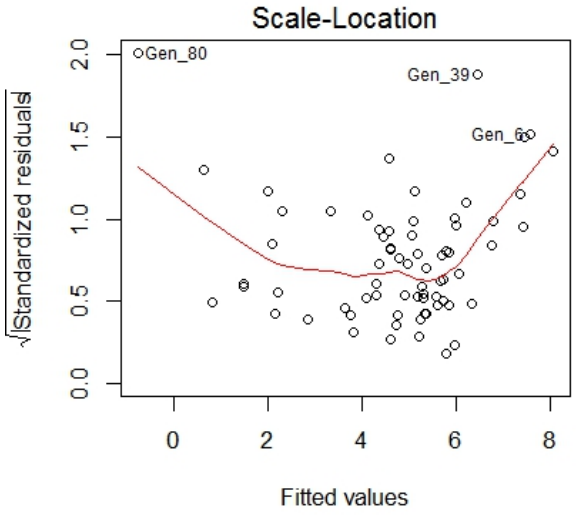
Proportion of sum of square explained by the 11 traits

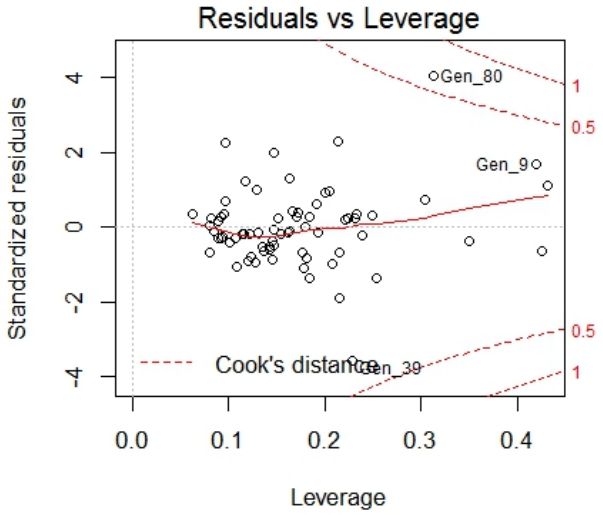
Linear regression





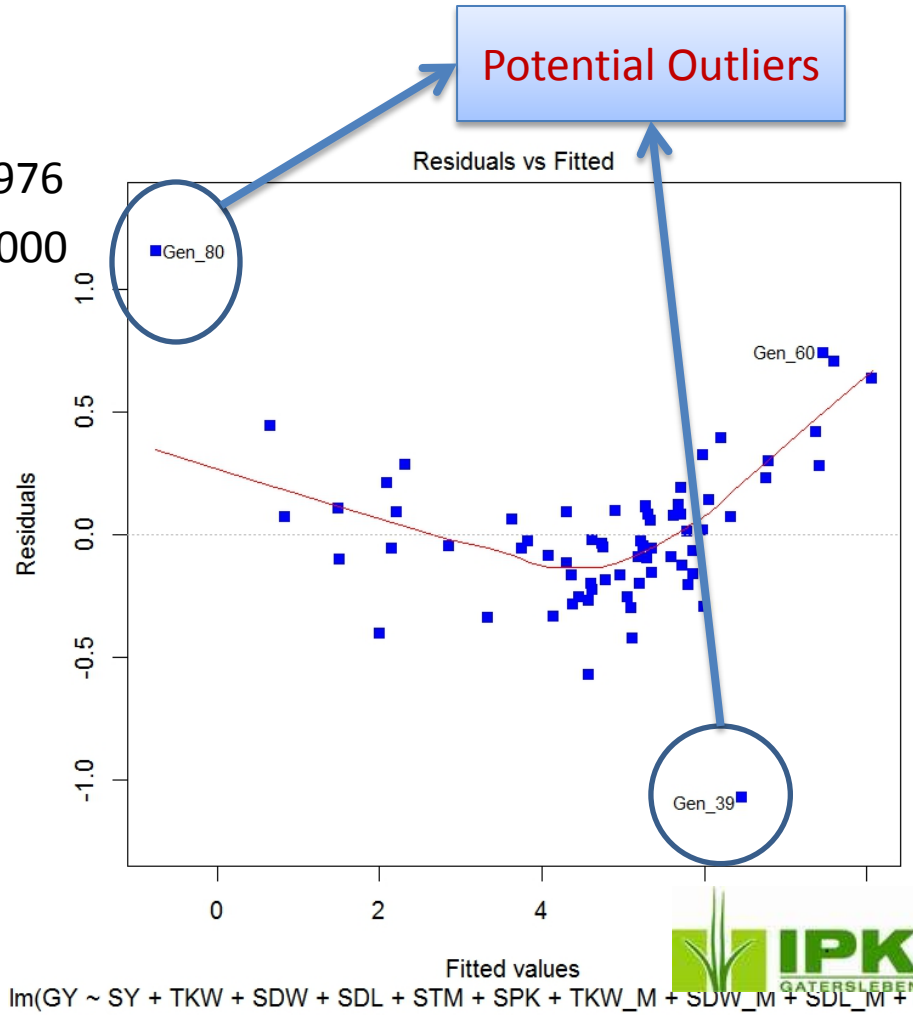

```
# diagnostic plots
layout(matrix(c(1:4),2,
2,byrow=TRUE))
plot(R_lm)
#dev.off()
```





Linear regression

```
plot(R_lm,which=1,col="blue",pch=15)
# Assessing Outliers, Bonferonni p-value for observations with most extreme residuals
outlierTest(R_lm)
rstudent unadjusted p-value Bonferonni p
Gen_80  4.747541       1.4282e-05   0.00099976
Gen_39 -3.978719       1.9799e-04   0.01386000
```



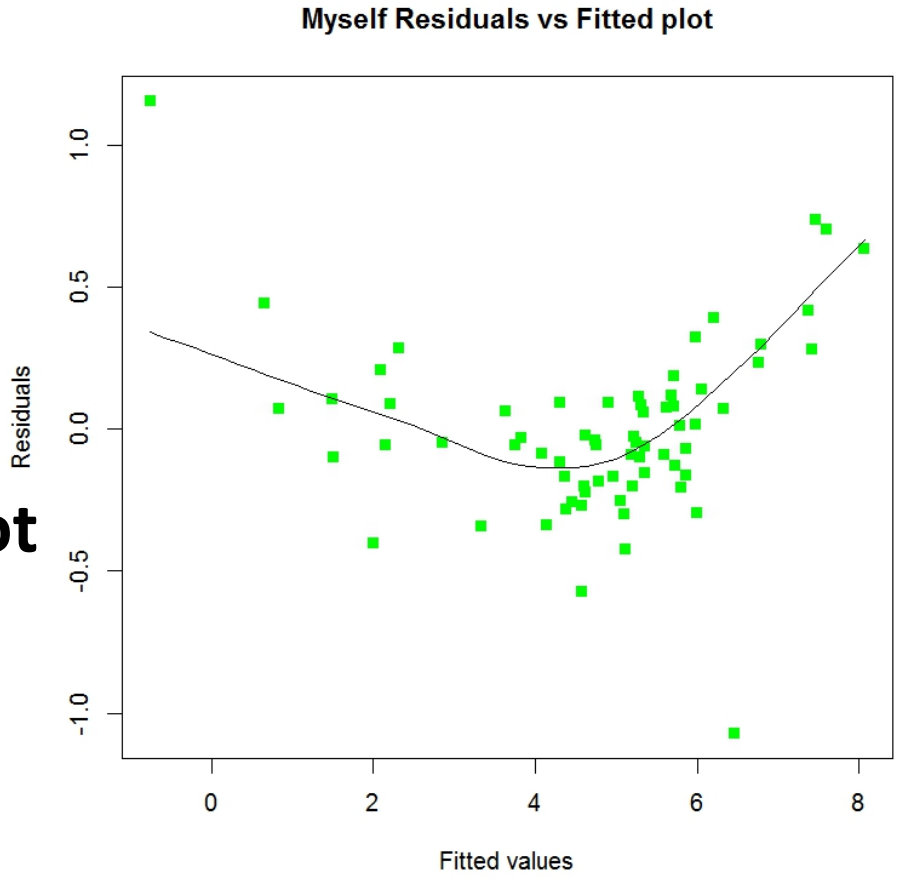
Residuals vs Fitted values

If points in this figure are not flat, then maybe there is a non-linear relationship between criterion variable and predictor variables

Linear regression

Residuals vs Fitted values

The lines in the middle is a smooth curve fitted by **locally weighted scatterplot smoothing**(loess), see R code bellow:



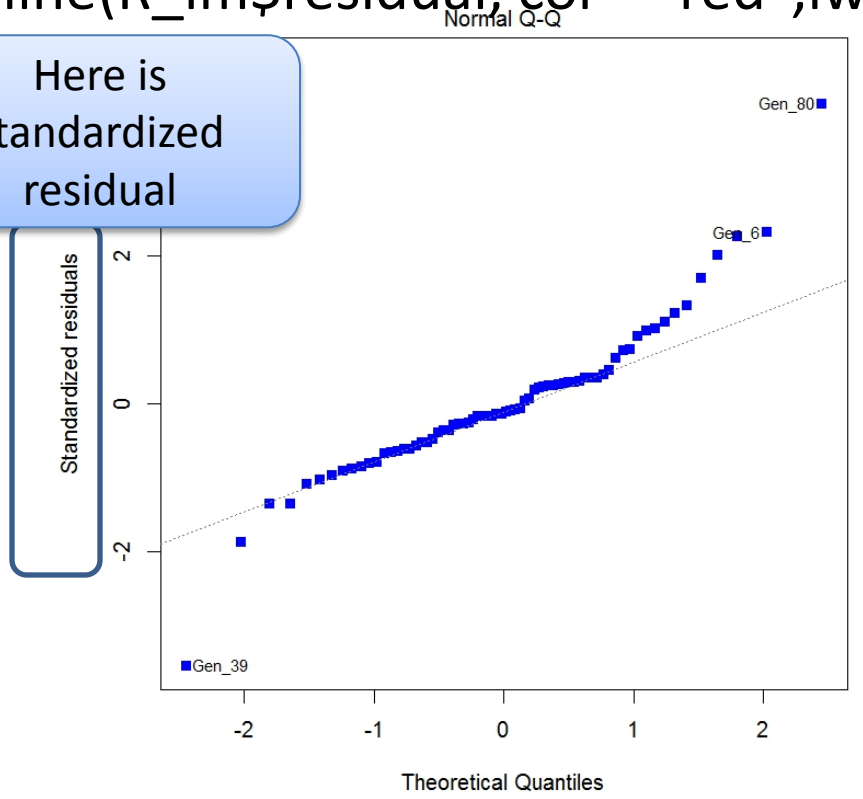

```
scatter.smooth(R_lm$residuals~R_lm$fitted.values , span = 2/3, degree = 1,pch=15,
col="Green",     ylab = "Residuals", xlab = "Fitted values", main = "Myself Residuals vs
Fitted plot")
```

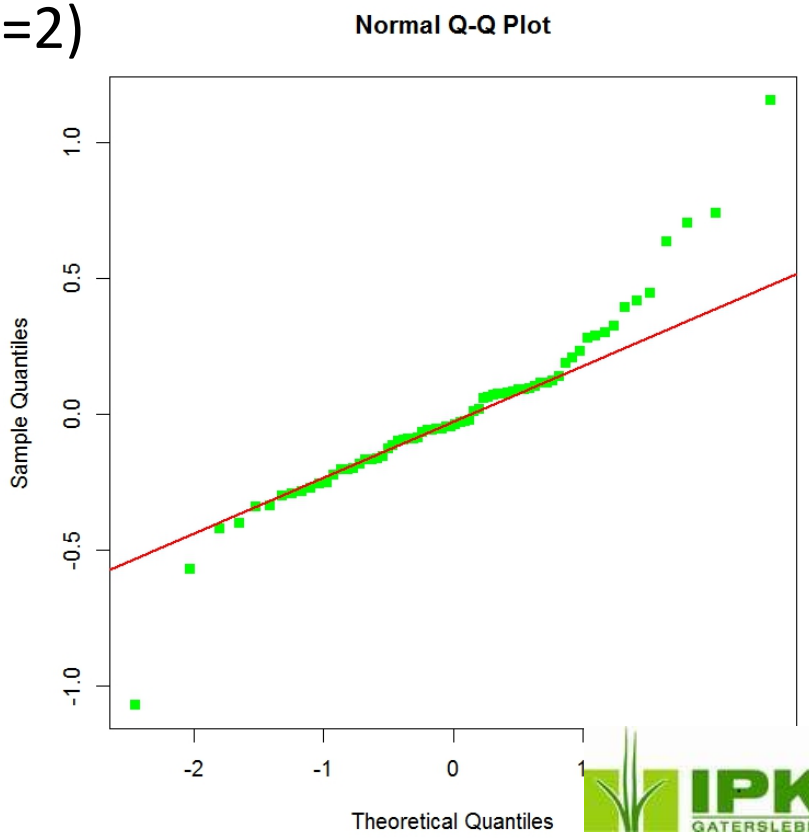
Linear regression

QQ-plot for residuals

```
plot(R_lm,which=2,col="blue",pch=15)
qqnorm(R_lm$residual,col="green",pch=15);
qqline(R_lm$residual, col = "red",lwd=2)
```

Here is standardized residual





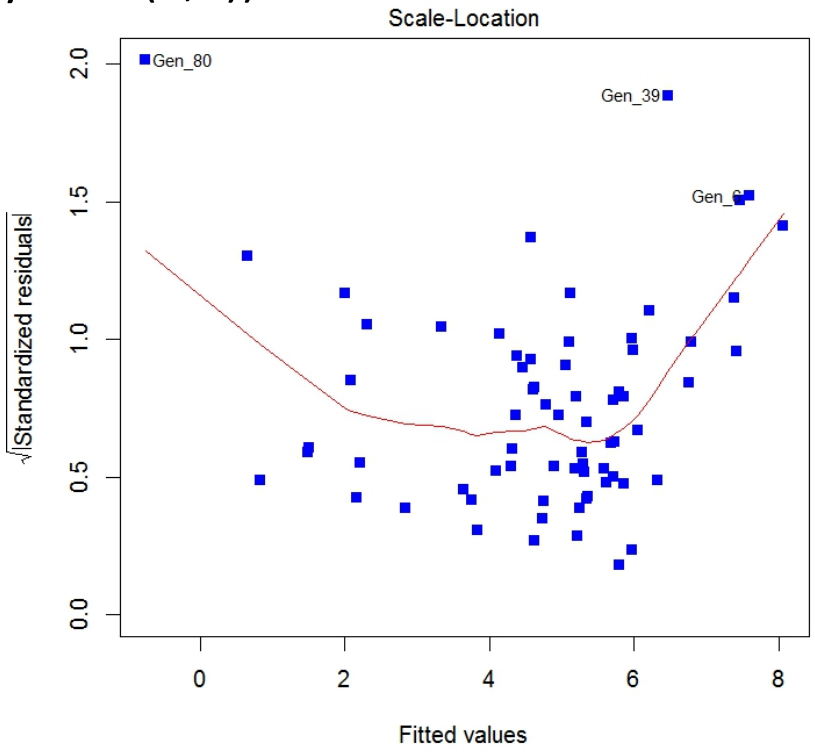
Linear regression

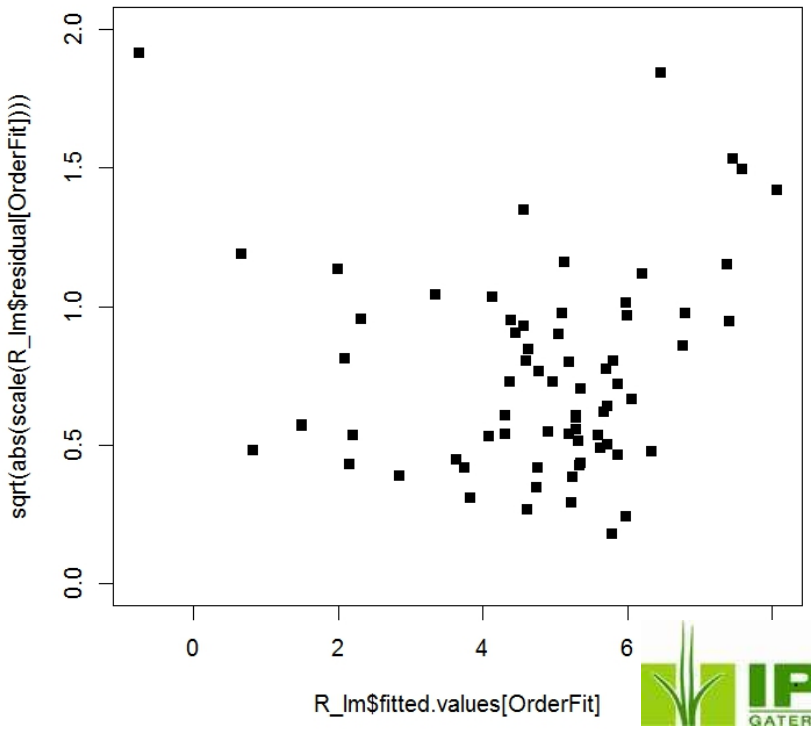
Scale-Location plot

layout(matrix(c(1:2),1,2))

plot(R_lm,which=3,col="blue",pch=15) OrderFit<-order(R_lm$fitted.values)

plot(R_lm$fitted.values[OrderFit],sqrt(abs(scale(R_lm$residual[OrderFit]))),pch=15,ylim=c(0,2))





Linear regression

Residuals vs Leverage

Find influential data points

When points are outside of the Cook's distance (the red dash lines) then they are influential to the regression results.

> lm(Data[-80,-1])

Coefficients:

(Intercept)

-4.649593

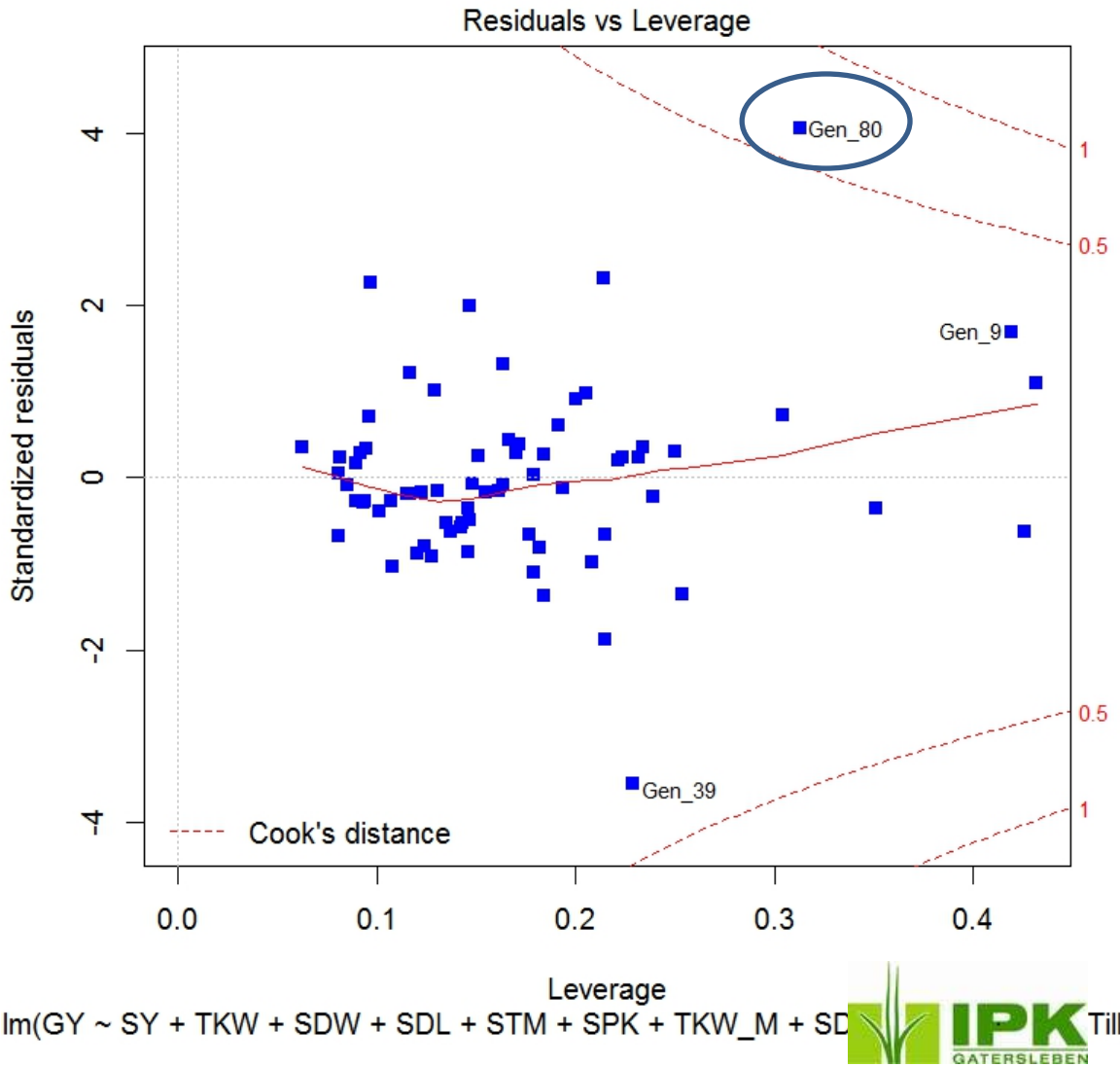
> lm(Data[,-1])

Coefficients:

(Intercept)

-3.6565105

Very large changes!



Limit of linear regression

Number of observations must be larger than number of variables, otherwise, linear regression will not work.

For example, if we have only 10 genotypes in the data

```
> lm(Data[1:10,-1])

Call:
lm(formula = Data[1:10, -1])

Coefficients:
(Intercept)         SY        TKW        SDW        SDL        STM        SPK
   41.28848    0.01516   -0.14944    0.04158    1.50230    0.11337   -2.74136
     TKW_M      SDW_M      SDL_M     Tiller       Leaf
   -0.44513    0.64289         NA         NA         NA
```

Analysis of variance (ANOVA)

Example Data 2: We have 27 wheat genotypes treated with three different levels of fertilizer(Lower, higher and normal) in green house.

	Genotype	Treatment	GY	SY	TKW	...	Leaf
1	Line_1	Control	0.8	57	14	...	4
27	Line_1	Higher	1.1	102	10.8	...	5
54	Line_1	Lower	0.6	39	15.4	...	3
19	Line_10	Control	6.4	142	45.1	...	2

Question 1: Are there significant differences for each traits between different treatments?

Analysis of variance (ANOVA)

One-way ANOVA

One-way ANOVA is a kind of generalization of a two-sample t-test which compares the means of several groups simultaneously.

In one-way ANOVA, there is one dependent variable and one categorical variable (for example treatments)

Analysis of variance (ANOVA)

Question 1: Are there significant differences for each trait between different treatments?

```
> # Analysis of variance(ANOVA)
> Data2<-read.table("Data treatments.txt",header=TRUE)
> Anova_GY<-aov(GY~Treatment,Data2)
> Anova_GY
Call:
   aov(formula = GY ~ Treatment, data = Data2)
Terms:
                    Treatment Residuals
Sum of Squares    109.752   446.012
Deg. of Freedom         2        68
Residual standard error: 2.561054
Estimated effects may be unbalanced
9 observations deleted due to missingness
```

Analysis of variance (ANOVA)

Question 1: Are there significant difference for each trait between different treatments?

```
> # Analysis of variance (ANOVA) for one trait
> summary(Anova_GY)
             Df Sum Sq Mean Sq F value   Pr(>F)
Treatment     2  109.8   54.88   8.367 0.000564 ***
Residuals    68  446.0    6.56
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
9 observations are deleted due to missingness
```

P-value<0.05, there is significant difference between treatments

Treatment	Mean	Max	Min	Number of genotypes
Control	4.652	11.6	0.8	26
Higher	6.348	11.2	1.1	27
Lower	3.335	7.7	0.6	27

Analysis of variance (ANOVA)

```
> Traits<-colnames(Data2)[-c(1:3)]
> Pv<-NULL
> for(i in 1:8){
+ Anova_T<-aov(as.formula(paste(Traits[i],"~Treatment")),Data2)
+ Pv<-
rbind(Pv,cbind(Traits[i],as.numeric(summary(Anova_T)[[1]][["Pr(>F)"]][1]))
)
+ }
> Pv
     [,1]     [,2]
[1,] "GY"     "0.000564298417989619"
[2,] "SY"     "3.44916019343957e-06"
[3,] "TKW"    "0.565823910922791"
[4,] "SDW"    "0.160064612492598"
[5,] "SDL"    "0.350458299377 93"
[6,] "STM"    "1.65376818854514e-13"
[7,] "Tiller" "0.00315390132885176"
[8,] "Leaf"   "0.00423592925712039"
```

NO difference for TKW SDW and SDL traits

IPK GATERSLEBEN

Analysis of variance (ANOVA)

One-way ANOVA

ANOVA only showes statistically significant differences between groups, but it does not tell us which groups are different.

Post-hoc test can find out groups that differ from others after running ANOVA.

Example: Tukey's HSD (honest significant difference) test

Analysis of variance (ANOVA)

Example: Tukey's HSD (honest significant difference) test

```
> i<-1
> Anova_T<-aov(as.formula(paste(Traits[i],"~Treatment")),Data2)
> posthoc <- TukeyHSD(x=Anova_T, 'Treatment', conf.level=0.95)
> posthoc
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = as.formula(paste(Traits[i], "~Treatment")), data = Data2)
$Treatment
                    diff        lwr       upr     p adj
Higher-Control  1.695826 -0.07716973  3.468822 0.0637533
Lower-Control  -1.317391 -3.12694758  0.492165 0.1963471
Lower-Higher   -3.013217 -4.78621320 -1.240222 0.0003608
```

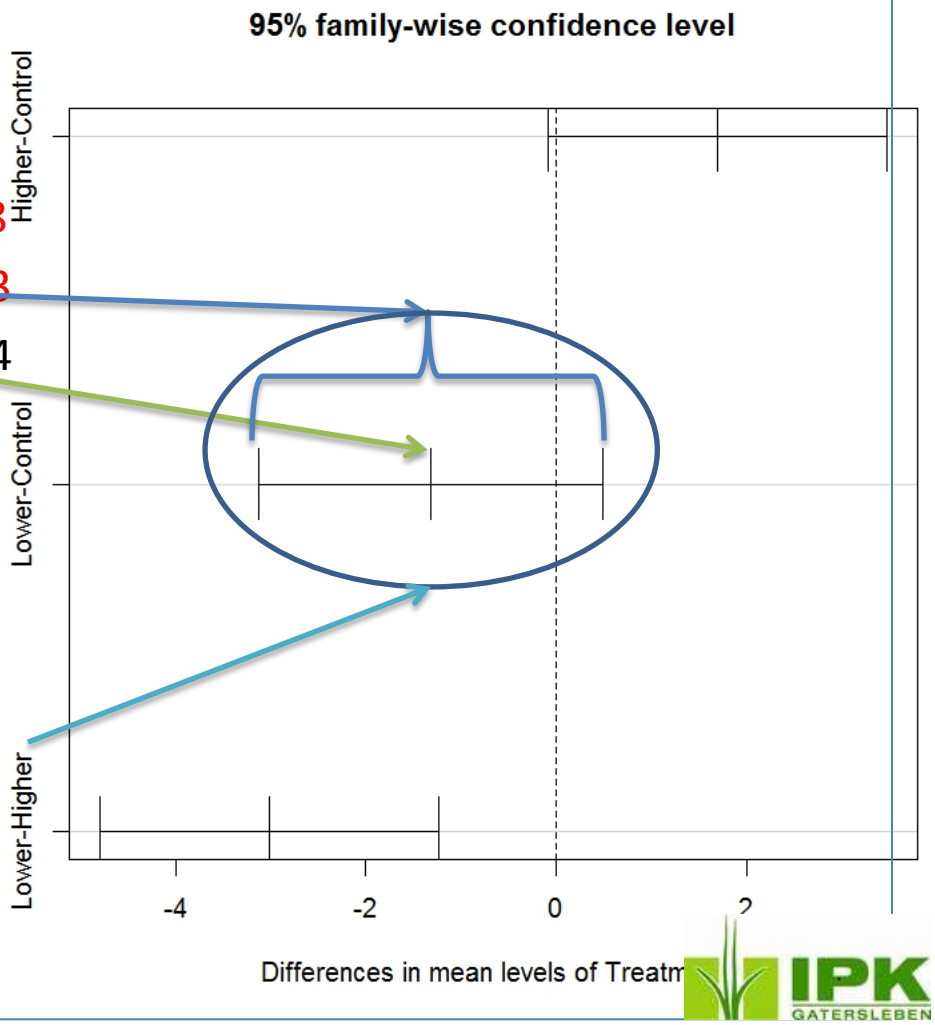
Analysis of variance (ANOVA)

To interpret Tukey's HSD test

> plot(posthoc)

> round(posthoc$Treatment,4)

	diff	lwr	upr	p adj
Higher-Control	1.6958	-0.0772	3.4688	0.0638
Lower-Control	-1.3174	-3.1269	0.4922	0.1963
Lower-Higher	-3.0132	-4.7862	-1.2402	0.0004



Confidence intervals for the difference in the means

Analysis of variance (ANOVA)

Two-way ANOVA

In two-way ANOVA, there is one dependent variable and two categorical variables (for example, country and treatment)

Two-way ANOVA is a extension of one way Anova, which tests the statistical significance of the two independent categorical variables and their interaction.

Example code (continue)

```
> Anova_Two<-
aov(as.formula(paste(Traits[i],"~Country+Treatment+Country*Treatment")),Data2
)
> summary(Anova_Two)
                Df Sum Sq Mean Sq F value   Pr(>F)
Country          2  377.5  188.75 163.228  < 2e-16 ***
Treatment        2   89.4   44.68  38.637 1.27e-11 ***
Country:Treatment  4   17.2    4.30   3.719   0.0089 **
Residuals       62   71.7    1.16
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
9 observations deleted due to missingness
>
```

Two variables “Country”, “Treatment” and their interaction “Country*Treatment” are all significantly not 0.

Anova and linear regression

```
> ## Anova and linear regression
> Lm_1<-lm(as.formula(paste(Traits[i],"~Treatment")),Data2)
> summary(Lm_1)
Call:
lm(formula = as.formula(paste(Traits[i], "~Treatment")), data = Data2)
Residuals:
    Min      1Q  Median      3Q     Max
-5.2480 -1.8935 -0.4348  1.6065  6.9478
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.6522     0.5340   8.712 1.12e-12 ***
TreatmentHigher   1.6958     0.7400   2.292   0.0250 *
TreatmentLower   -1.3174     0.7552  -1.744   0.0856 .
F-statistic: 8.367 on 2 and 68 DF,  p-value: 0.0005643
```

Same p-value to anova

Anova and linear regression

```
> ## Anova and linear regression
> anova(Lm_1)
Analysis of Variance Table

Response: GY
          Df Sum Sq Mean Sq F value    Pr(>F)
Treatment  2 109.75  54.876  8.3665 0.0005643 ***
Residuals 68 446.01   6.559
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Same as R function aov()

IPK GATERSLEBEN

Anova and linear regression

```
> Lm_2<-lm(as.formula(paste(Traits[i],"~Country+Treatment+Country*Treatment")),Data2)
> summary(Lm_2)
…….
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      8.6200     0.4809  17.924  < 2e-16 ***
CountryFrance                   -3.8200     0.5890  -6.486 1.66e-08 ***
CountryGermany                  -6.6325     0.6130 -10.819 6.53e-16 ***
TreatmentHigher                  1.3086     0.6297   2.078 0.041836 *
TreatmentLower                  -2.7800     0.6801  -4.088 0.000128 ***
CountryFrance:TreatmentHigher    0.4814     0.7923   0.608 0.545652
CountryGermany:TreatmentHigher  -0.3836     0.8280  -0.463 0.644803
CountryFrance:TreatmentLower     1.4000     0.8330   1.681 0.097844 .
CountryGermany:TreatmentLower    2.4550     0.8670   2.832 0.006236 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

p-value for each effect

IPK GATERSLEBEN

Anova and linear regression

```
> ## Anova and linear regression
> anova(Lm_2)
Analysis of Variance Table

Response: GY
                    Df Sum Sq Mean Sq F value    Pr(>F)
Country              2 377.51 188.753 163.228 < 2.2e-16 ***
Treatment            2  89.36  44.680  38.637 1.271e-11 ***
Country:Treatment    4  17.20   4.301   3.719  0.008905 **
Residuals           62  71.70   1.156
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Same as R function aov()

IPK GATERSLEBEN

Notes on ANOVA

- ANOVA is based on the assumption that data is balanced desigined, normally distributed , and **categorical variables** are independent

- ANOVA can produce same tests and results when compared to linear regression.

- If data is not balanced designed, then a mixed-model approach is normally required to do analyses.

Thank you for your interest!

Pictures: Frank Schröder